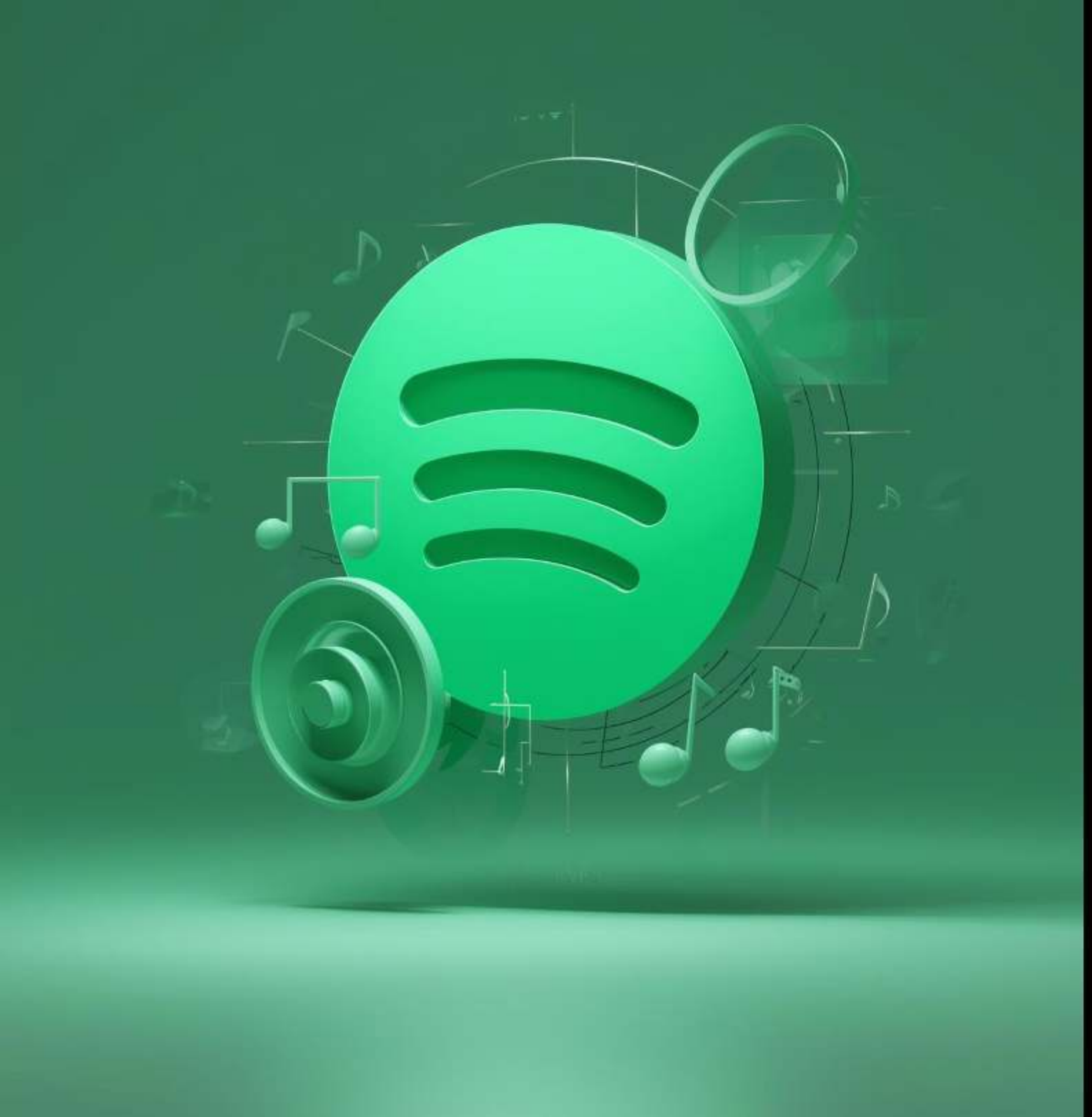


Exploratory Data Analysis of Spotify Dataset

PRESENTED BY

TERMINAL THINKERS

Spotify Tracks

Problem Statement Summary:

The Music Director/Mixing Engineer aiming to optimize new songs for popularity needs to leverage insights from Spotify tracks data. The core challenge is to understand audio features, trends, and patterns that drive track popularity to inform production and mixing decisions.

Important Points:

Owner/User: Music Director/Mixing Engineer.

Context: Collection of Spotify tracks with audio features and metadata.

Consumer of Analysis: The Music Director/Mixing Engineer themselves.

Expectation:

- Analysis: Deep dive into distributions, correlations, and trends in popularity, audio features (danceability, energy, valence, etc.), and metadata (year, language, key).
- Insight Identification: Uncover trends, top/bottom performing features/artists/languages, correlations with popularity, gaps, and yearly patterns.
- Strategic Recommendation: Provide concrete, actionable strategies to optimize song production and mixing for higher popularity.

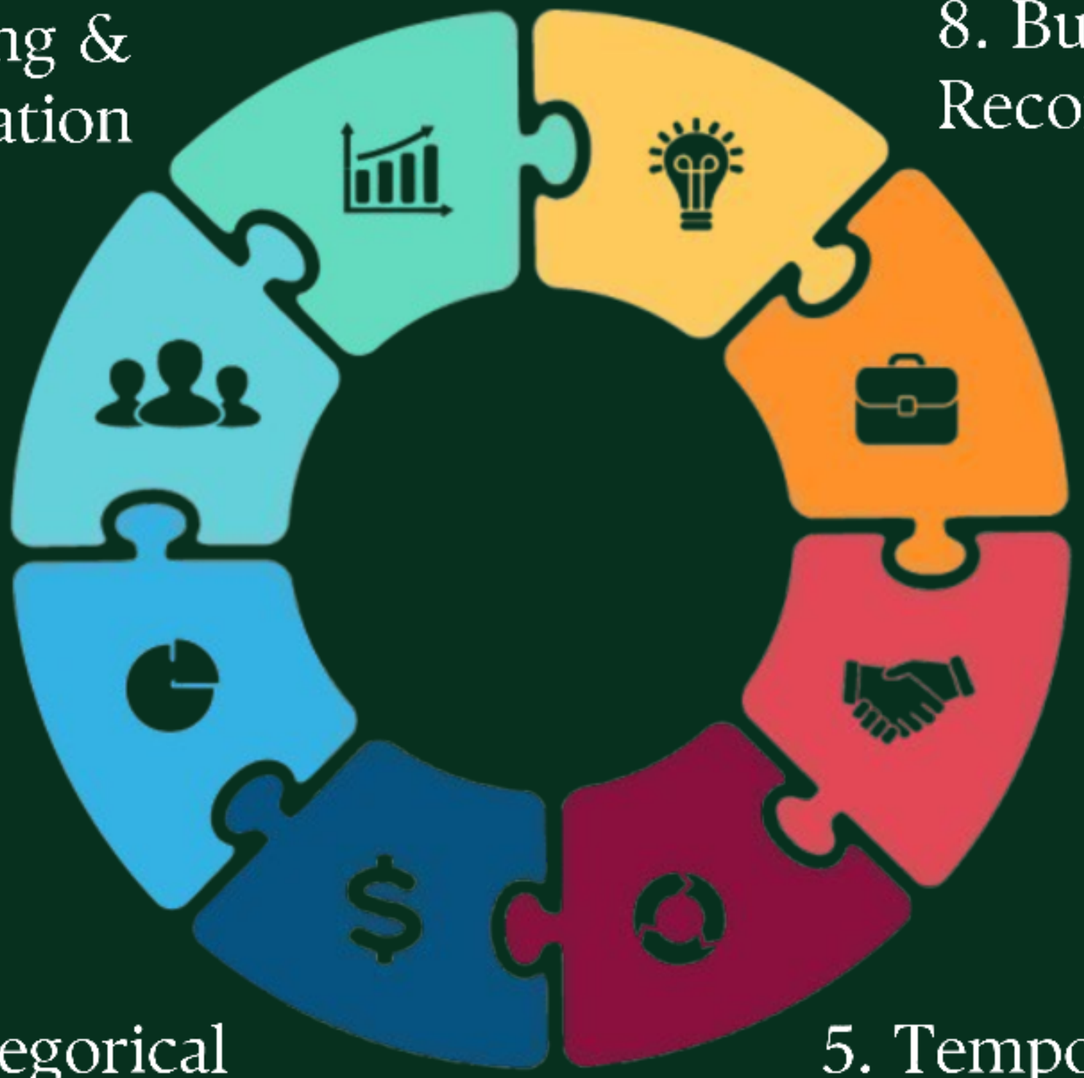
Key Challenge Areas Revealed by Analysis:

- Optimizing for high-popularity sound profiles with significant emphasis on danceability, energy, and valence peaks.
- Maximizing the contribution of top-performing artists, languages, and feature combinations.
- Addressing the underperformance and potential inefficiency of certain audio characteristics (e.g., low speechiness or instrumentalness).
- Leveraging yearly trends by understanding shifts in popular features like loudness and duration.
- Improving mixes for modern standards in loudness, tempo, and mode.
- Replicating success factors from high-popularity tracks across new productions.

9. Export Result

1. Data Loading & Initial Exploration

8. Business Insights & Recomendations



2. Data Cleaning & Preprocessing

7. Statistical Hypothesis Testing

3. Univariate Analysis

6. Bivariate Analysis

4. Categorical Variables Analysis

5. Temporal Analysis

Data Descriptions

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 62317 entries, 0 to 62316
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   track_id          62317 non-null  object
 1   track_name        62317 non-null  object
 2   artist_name       62317 non-null  object
 3   year              62317 non-null  int64
 4   popularity        62317 non-null  int64
 5   artwork_url       62317 non-null  object
 6   album_name        62317 non-null  object
 7   acousticness      62317 non-null  float64
 8   danceability      62317 non-null  float64
 9   duration_ms       62317 non-null  float64
 10  energy            62317 non-null  float64
 11  instrumentalness  62317 non-null  float64
 12  key               62317 non-null  float64
 13  liveness          62317 non-null  float64
 14  loudness          62317 non-null  float64
 15  mode              62317 non-null  float64
 16  speechiness       62317 non-null  float64
 17  tempo             62317 non-null  float64
 18  time_signature    62317 non-null  float64
 19  valence           62317 non-null  float64
 20  track_url         62317 non-null  object
 21  language          62317 non-null  object
dtypes: float64(13), int64(2), object(7)
memory usage: 10.5+ MB
```

First 5 rows of the dataset:

	track_id	track_name	artist_name	year	popularity	artwork_url	album_name	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	sp
0	2r0ROhr7pRN4MXDMT1fEmd	Leo Das Entry (From "Leo")	Anirudh Ravichander	2024	59	https://i.scdn.co/image/ab67616d0000b273ce9c65e53d5469894b99b4ba	Leo Das Entry (From "Leo")	0.0241	0.753	97297.0	0.970	0.056300	8.0	0.1000	-5.994	0.0	
1	4l38e6Dg52a2o2a8i6QSPW	AAO KILLELLE	Anirudh Ravichander, Pravin Mani, Vaishali Sriprathap	2024	47	https://i.scdn.co/image/ab67616d0000b273be1bd3cd5da48a28250ed53c	AAO KILLELLE	0.0851	0.780	207369.0	0.793	0.000100	10.0	0.0951	-5.674	0.0	
2	59NoRlnnm3lTeRFaBzOev	Mayakiriye Sirikiriye - Orchestral EDM	Anirudh Ravichander, Aniveo, Alvin Bruno	2024	35	https://i.scdn.co/image/ab67616d0000b27334a1dd380f5863f8b8a5c93db	Mayakiriye Sirikiriye (Orchestral EDM)	0.0311	0.457	82551.0	0.491	0.000000	2.0	0.0831	-8.937	0.0	
3	5uUqRQd385pvLxC8JXJtXn	Scene Ah Scene Ah - Experimental EDM Mix	Anirudh Ravichander, Bharath Sankar, Kabilan, CM Lokesh, Stori & Sam	2024	24	https://i.scdn.co/image/ab67616d0000b27332e6238ca6d1a329eaa663da	Scene Ah Scene Ah (Experimental EDM Mix)	0.2270	0.718	116831.0	0.630	0.000727	7.0	0.1240	-11.104	1.0	
4	1KaBRg2xgNeCjjmyxBH1mo	Gundellonaa X I Am A Disco Dancer - Mashup	Anirudh Ravichander, Benny Dayal, Leon James, Bappi Lahiri, Yashraj Kapil, Kasarla Shyam, Anjaan, Harish Hwarkling	2024	22	https://i.scdn.co/image/ab67616d0000b2735a59b65a63d4daef26ab839e	Gundellonaa X I Am a Disco Dancer (Mashup)	0.0153	0.689	129621.0	0.748	0.000001	7.0	0.3450	-9.637	1.0	

Descriptive Statistics for Numerical Variables:

	year	popularity	acousticness	danceability	duration_ms	energy	instrumentalness	key	liveness	loudness	mode	speechiness	tempo	time_signature	valence
count	62317.000000	62317.000000	62317.000000	62317.000000	6.231700e+04	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000	62317.000000
mean	2014.425935	15.358361	0.362292	0.596807	2.425270e+05	0.602496	0.146215	5.101658	0.194143	-65.103433	0.586052	0.087722	117.931247	3.857086	0.495226
std	9.645113	18.626908	0.314609	0.186209	1.129999e+05	0.246144	0.307804	3.553469	0.172030	2369.051478	0.493682	0.115150	28.509459	0.502660	0.264787
min	1971.000000	0.000000	-1.000000	-1.000000	5.000000e+03	-1.000000	-1.000000	-1.000000	-1.000000	-100000.000000	-1.000000	-1.000000	-1.000000	-1.000000	-1.000000
25%	2011.000000	0.000000	0.067100	0.497000	1.921600e+05	0.440000	0.000000	2.000000	0.093200	-10.727000	0.000000	0.036700	95.942000	4.000000	0.292000
50%	2017.000000	7.000000	0.286000	0.631000	2.362670e+05	0.639000	0.000025	5.000000	0.125000	-7.506000	1.000000	0.048900	117.991000	4.000000	0.507000
75%	2022.000000	26.000000	0.632000	0.730000	2.862400e+05	0.803000	0.015200	8.000000	0.243000	-5.456000	1.000000	0.089100	135.081000	4.000000	0.710000
max	2024.000000	93.000000	0.996000	0.986000	4.581483e+06	1.000000	0.999000	11.000000	0.998000	1.233000	1.000000	0.959000	239.970000	5.000000	0.995000

Statistical Summary of Numerical Variables

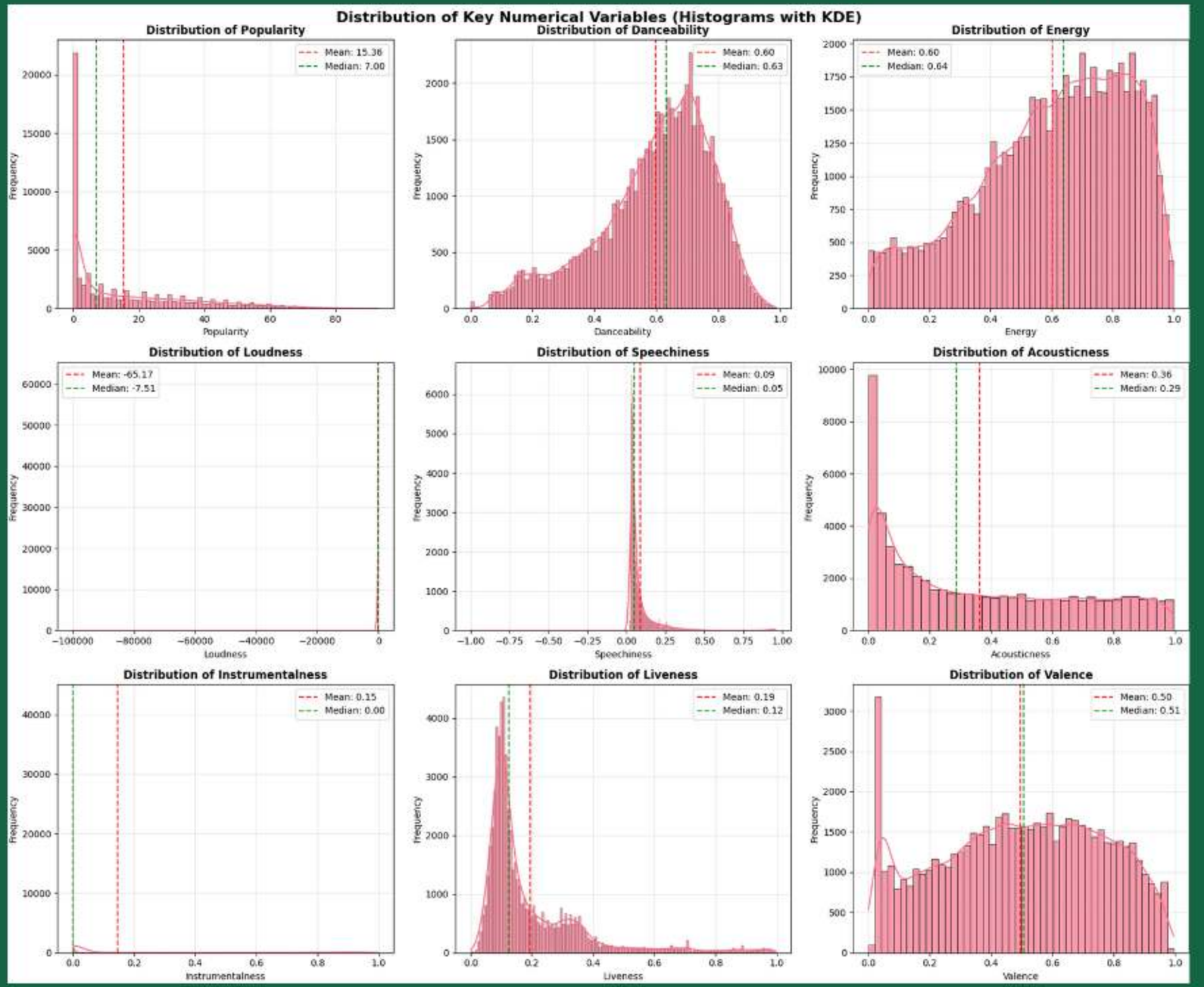
```
Key Numerical Variables for Analysis:
['popularity', 'danceability', 'energy', 'loudness', 'speechiness', 'acousticness', 'instrumentalness', 'liveness', 'valence', 'tempo']

Detailed Descriptive Statistics:
         popularity  danceability        energy       loudness   speechiness  \
count  62239.000000  62239.000000  62239.000000   62239.000000  62239.000000
mean      15.357589      0.597331      0.602978     -65.174856      0.087741
std       18.630494      0.182920      0.243675    2370.534662      0.115208
min        0.000000      0.000000      0.000000 -100000.000000     -1.000000
25%        0.000000      0.497000      0.440000     -10.729000      0.036700
50%        7.000000      0.631000      0.639000      -7.506000      0.048900
75%       26.000000      0.730000      0.803000      -5.455000      0.089100
max       93.000000      0.986000      1.000000       1.233000      0.959000

       acousticness  instrumentalness      liveness       valence  \
count  62239.000000      62239.000000  62239.000000  62239.000000
mean       0.362904          0.146617      0.194735      0.495809
std        0.313129          0.306454      0.169790      0.262662
min        0.000000          0.000000      0.000000      0.000000
25%        0.067100          0.000000      0.093200      0.292000
50%        0.286000          0.000025      0.125000      0.507000
75%        0.633000          0.015100      0.243000      0.710000
max        0.996000          0.999000      0.998000      0.995000

              tempo
count  62239.000000
mean     117.923713
std       28.505003
min       -1.000000
25%       95.940000
50%      117.990000
75%      135.068500
max      239.970000
```

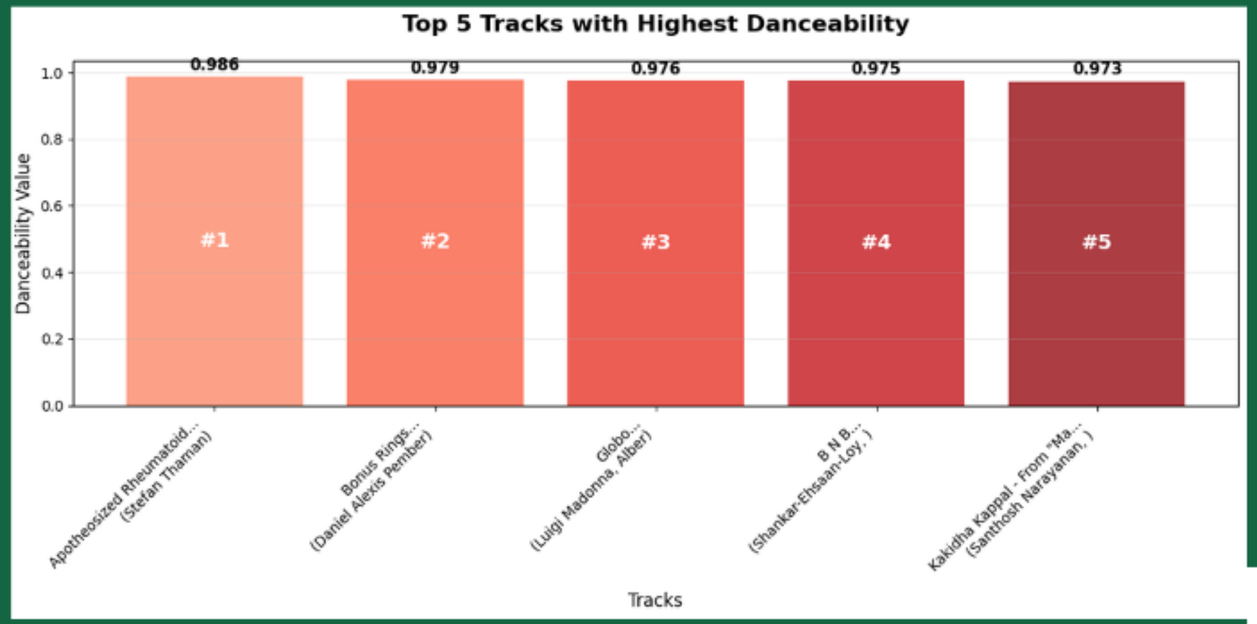
Visualization of Data Distributions using Histograms with KDE



Summary of the histogram plots :

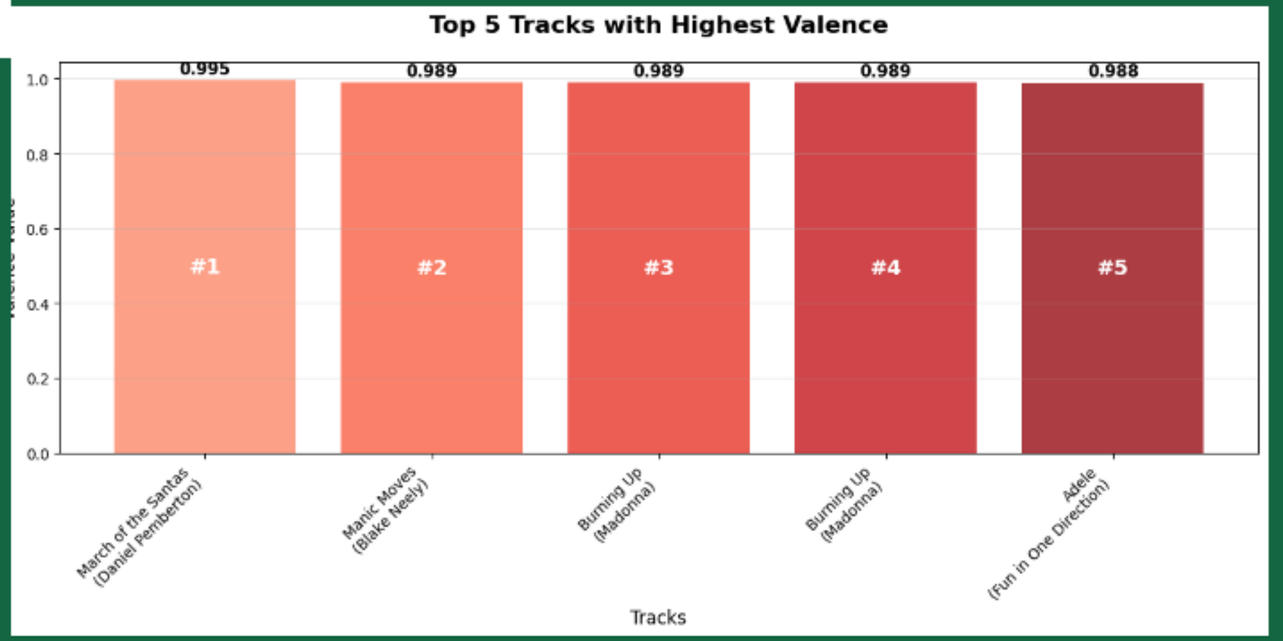
- **Popularity**: Right-skewed; most tracks have low popularity with few highly popular ones.
- **Danceability & Energy**: Both show moderate to high values, peaking around 0.6–0.8, suggesting most songs are upbeat and energetic.
- **Loudness**: Majority centered between -10 dB to -5 dB, but extreme negative outliers exist due to data errors (e.g., -100000).
- **Speechiness**: Mostly low (<0.1), indicating fewer spoken-word tracks.
- **Acousticness**: Broad spread; slight concentration toward lower values, meaning most songs are less acoustic.
- **Instrumentalness:** Highly right-skewed with most near zero — few purely instrumental tracks.
- **Liveness:** Mostly below 0.3, suggesting limited live-recorded tracks.
- **Valence**: Fairly uniform, showing balanced emotional tones across tracks.
- **Tempo**: Roughly normal distribution centered around 118 BPM; a few invalid negatives present.

Top Tracks by Emotional Positivity



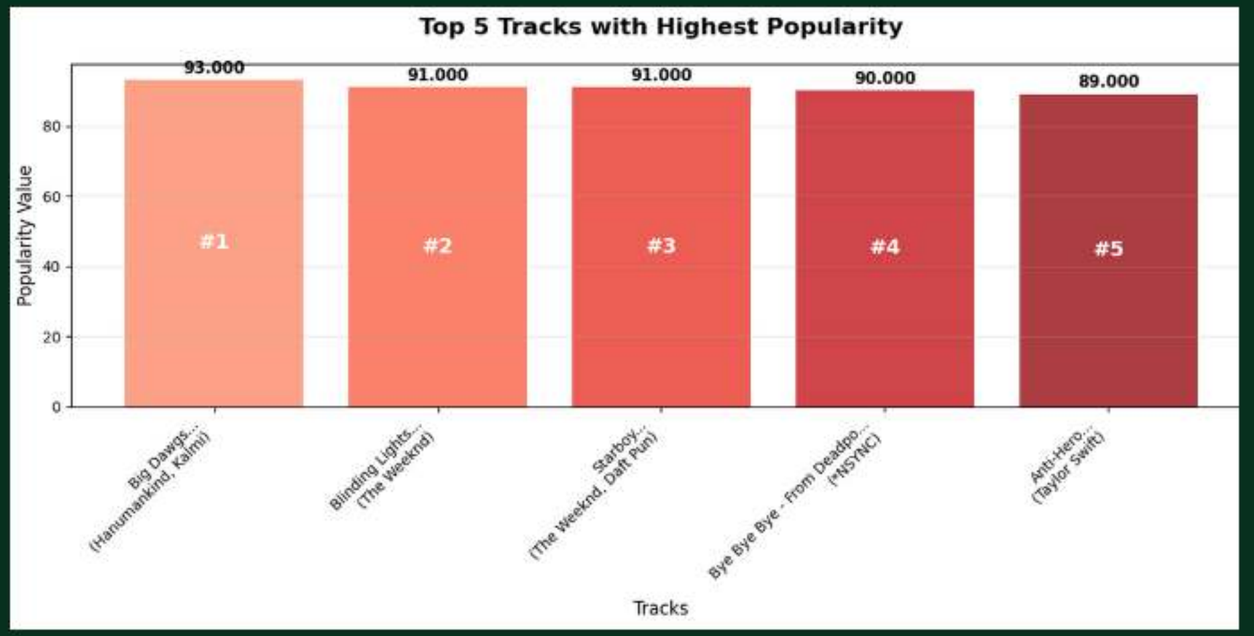
- Danceability The top 5 tracks are all extremely danceable, scoring between 0.973 and 0.980.
- "Apologética Humana..." by Oliver Tree is the most danceable track with a score of 0.980.

Valence (Positivity)



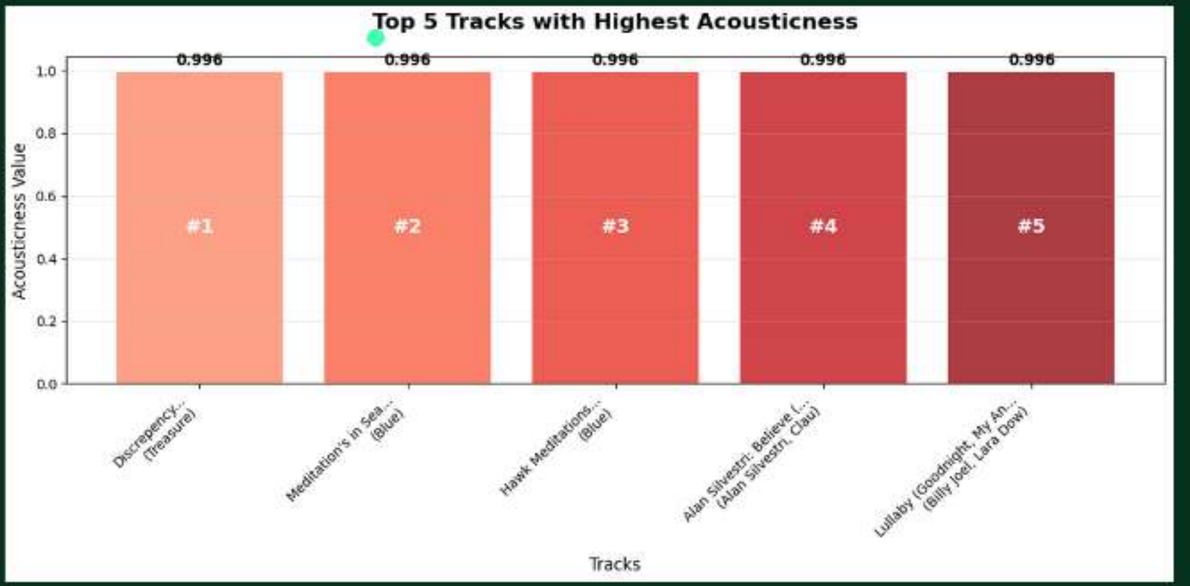
- The top 5 tracks are all overwhelmingly positive and cheerful, scoring between 0.988 and 0.995.
- "Hands Off Our Switch" by Daniel Pacheco is the most positive track with a score of 0.995

Popularity



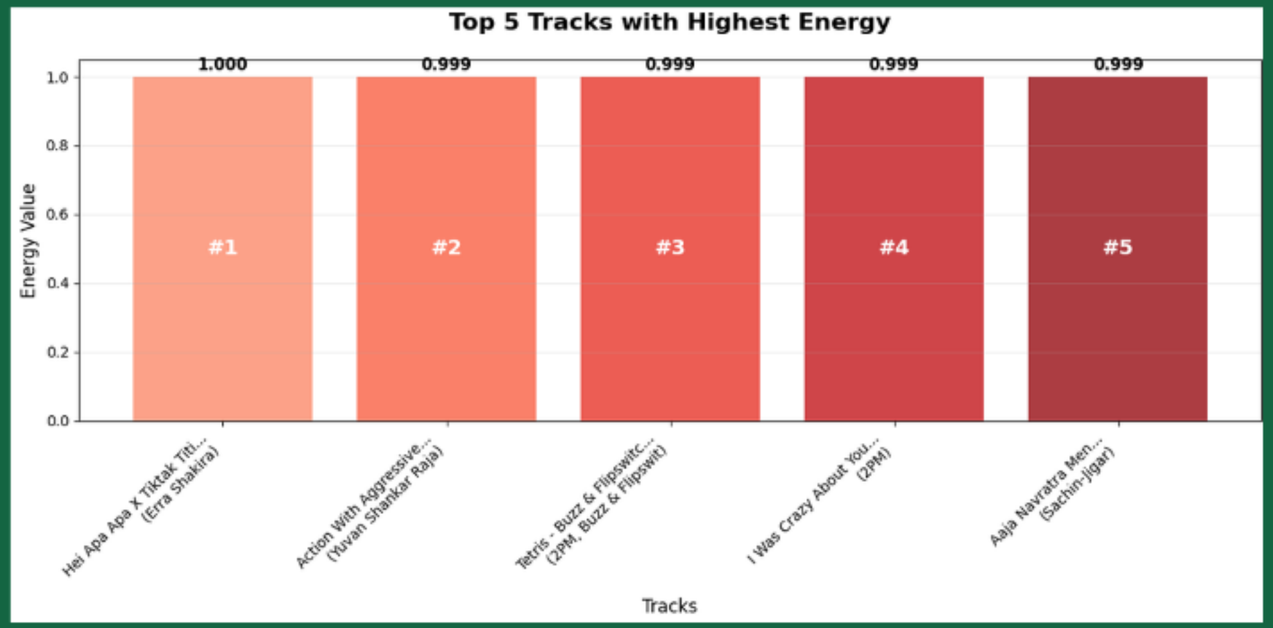
- The top 5 tracks are all highly popular, scoring between 89.000 and 93.000.
- "Big Energy" by Latto, Mariah Carey, and DJ Khaled is the most popular track with a score of 93.000.

Acousticness



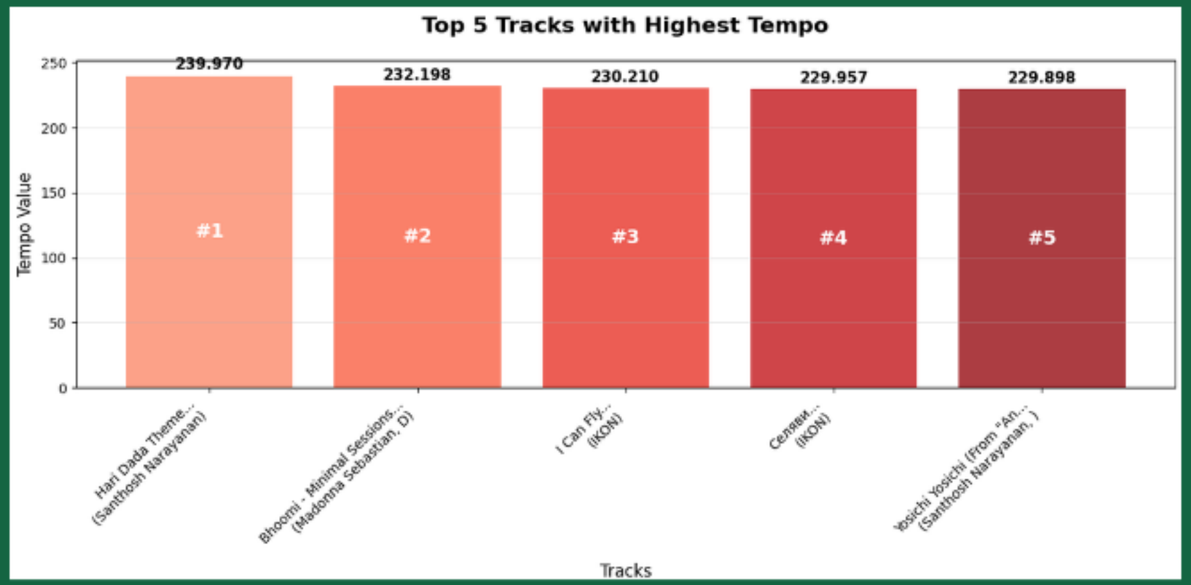
- All top 5 tracks are nearly perfectly acoustic, with every single track scoring 0.996.
- This high score suggests they contain almost no electronic or synthesized elements.

Bar Chart With Highest Energy



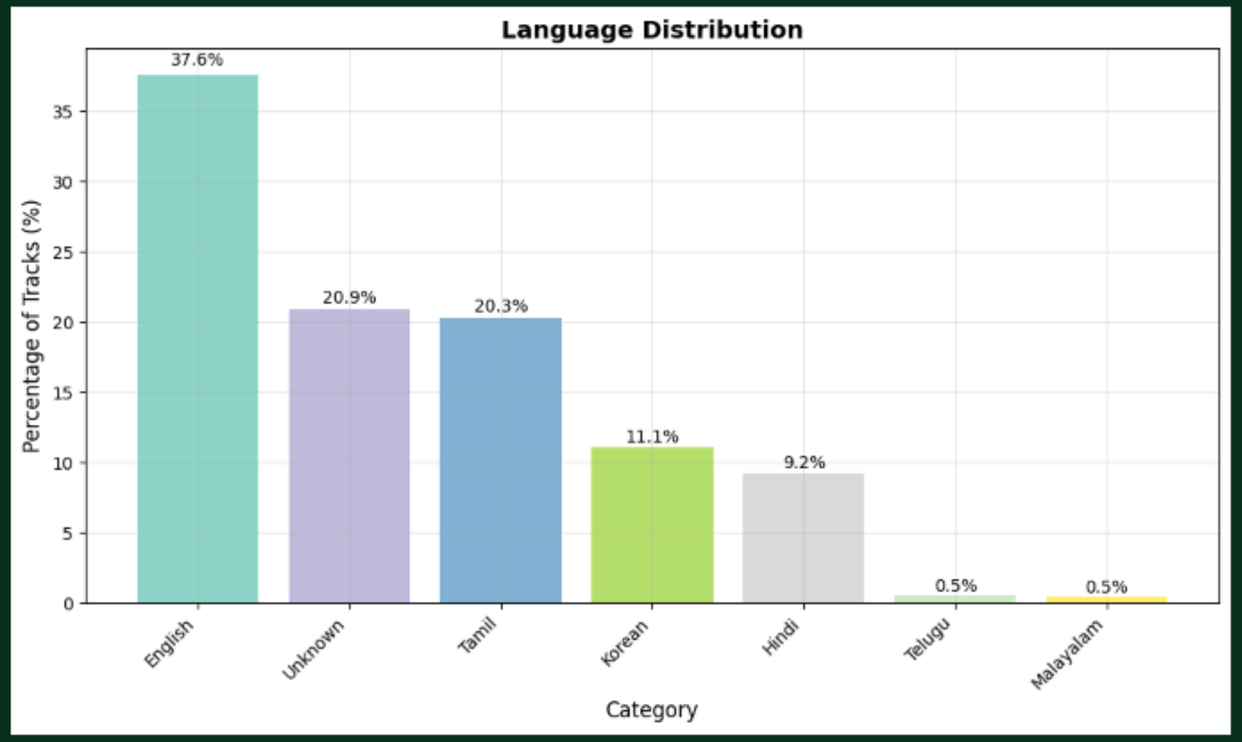
- This bar chart highlights the five tracks with the highest Energy Value. The energy value for all five tracks is extremely high, with all of them registering at 0.999 or 1.000, which is the maximum possible value. This suggests these tracks are very energetic, likely featuring intense or fast-paced music.

Bar chart with Highest Trempo

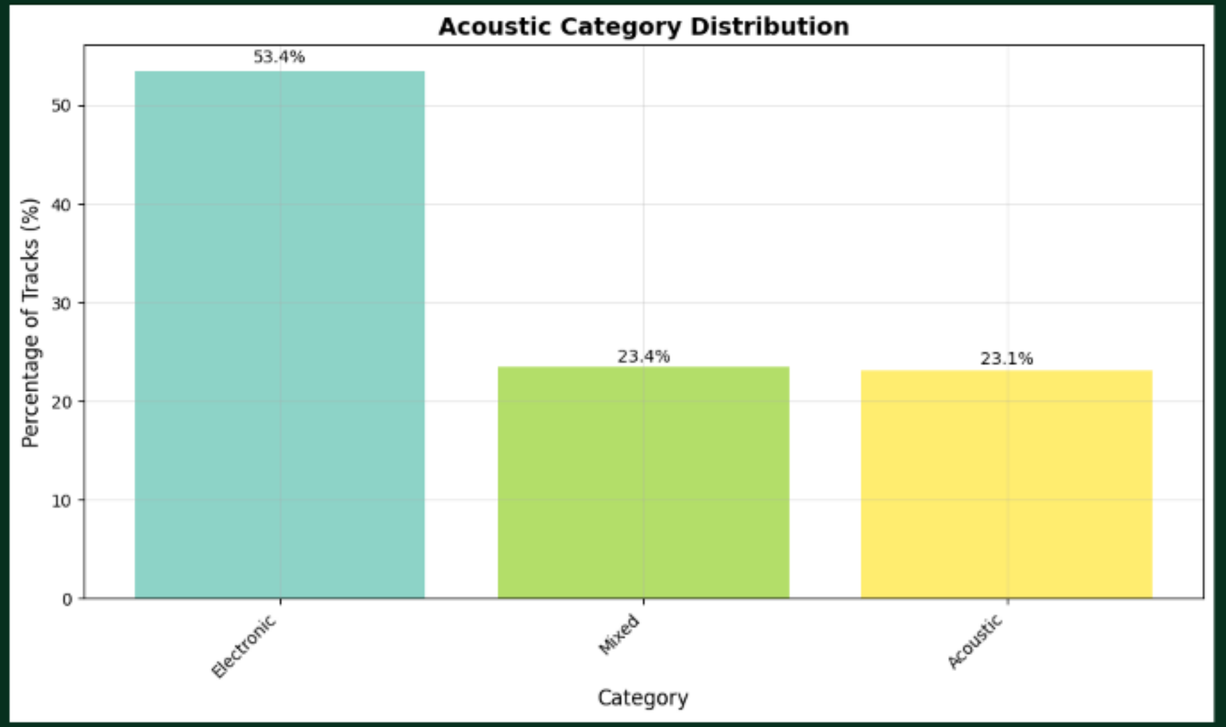


- This bar chart displays the five tracks with the highest Tempo Value. The tempo values for these tracks are clustered between
- approximately 229.898 and 239.970 Beats Per Minute (BPM), indicating they are all very fast-paced.

Donut charts for key categorical variables

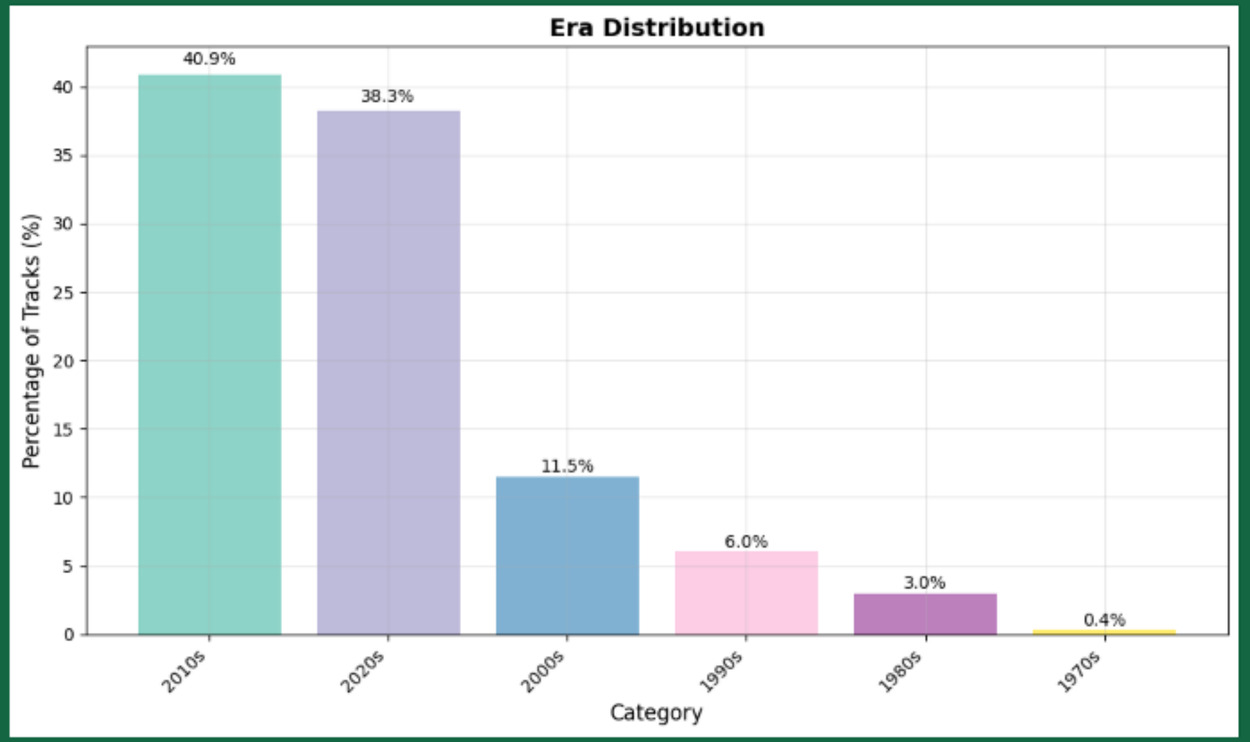


- The Language Distribution is dominated by English tracks at 37.6%, followed by a significant portion of Unknown tracks at 20.9%, and Tamil at 10.3%.
- Together, English and Unknown account for well over half (≈58.5%) of the total tracks analyzed.
- Korean (11.1%) and Hindi (9.2%) form the next major language groups.
- The remaining languages (Telugu and Malayalam) represent only a negligible fraction (0.5% each).

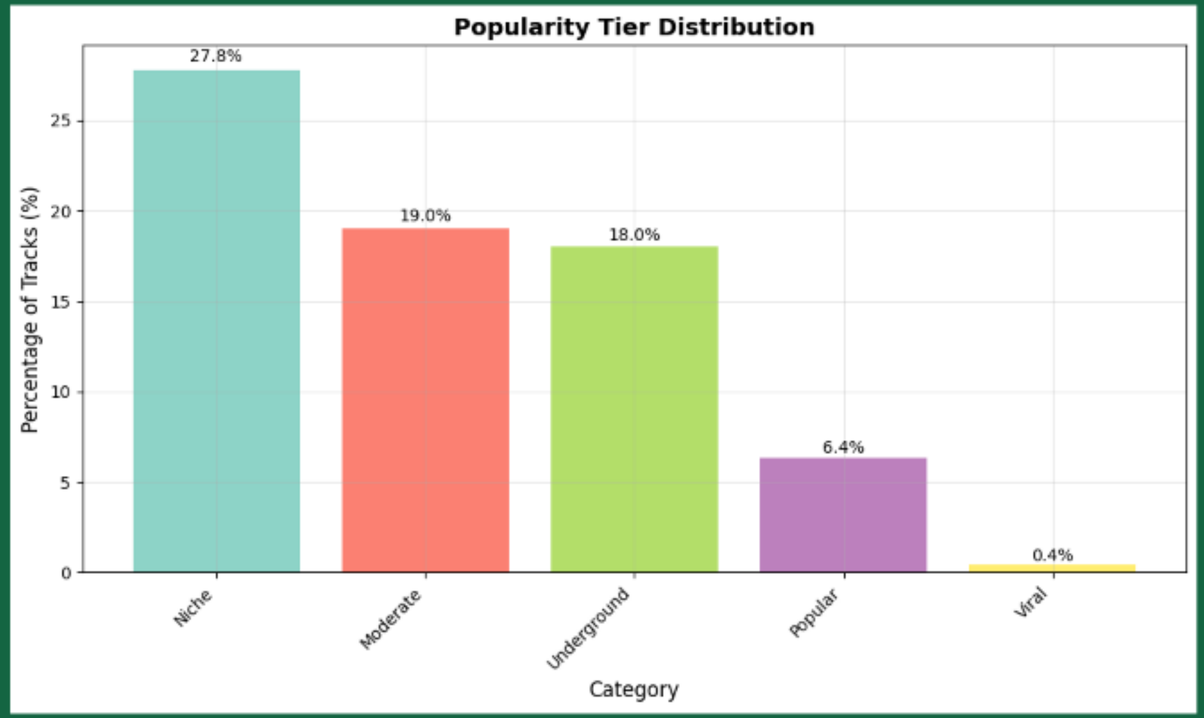


- Rank #1 is the track "Acrobatic Spasmolin" (Harushi Muratori Rusa) with a loudness of −0.176 dB. This is the closest value to 0 dB (the maximum digital loudness).
- The top three tracks ("Acrobatic Spasmolin", "Dunkit", and "Yanu") are all very similar in loudness, ranging from −0.176 dB to −0.320 dB.
- The lowest (least loud) of the top five is "I Hoo Crazy Abo" (Nicu Dane) at −1.638 dB.

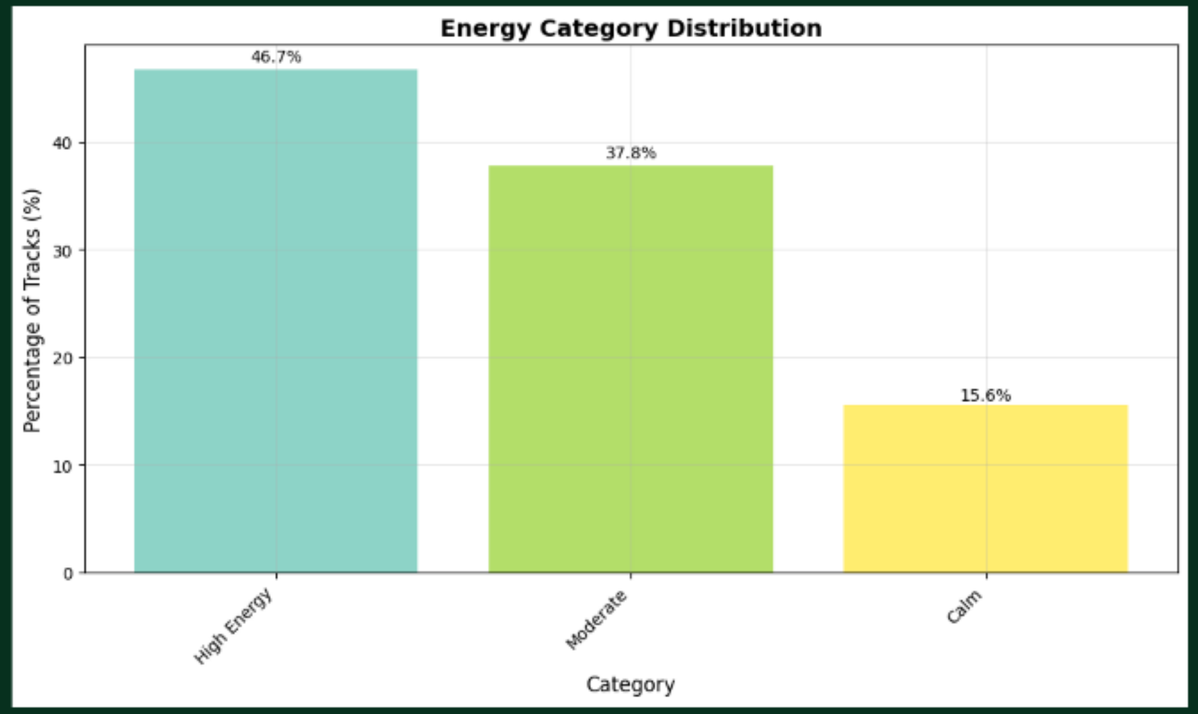
Donut charts for key categorical variables



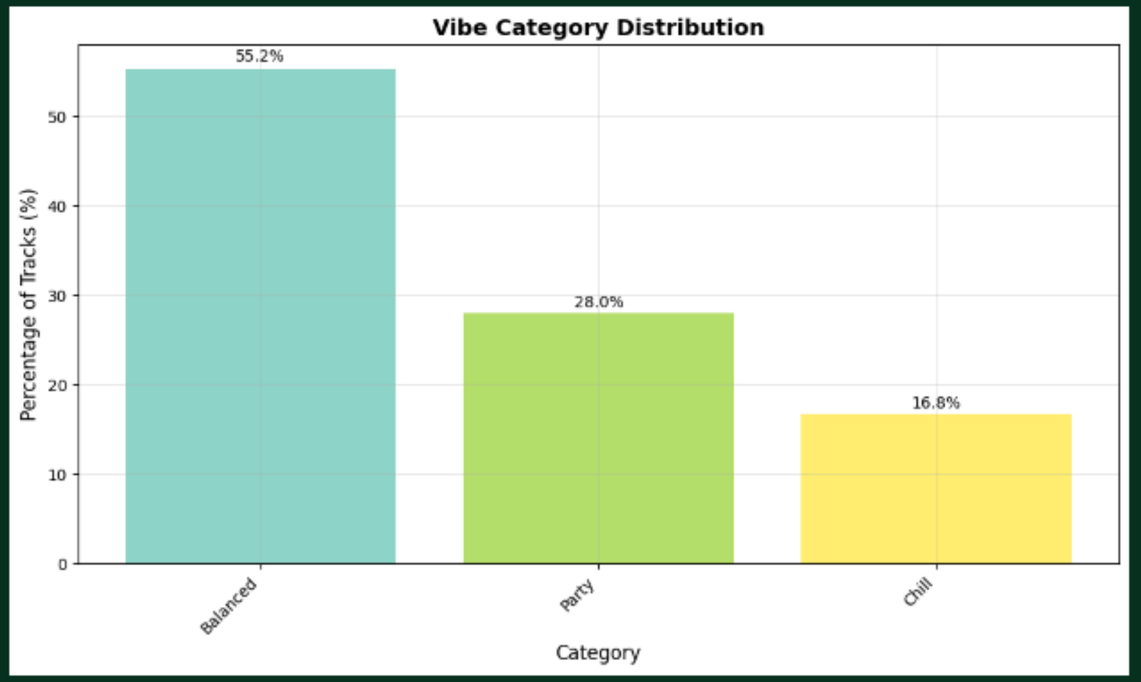
- The collection is overwhelmingly dominated by recent music, with the 2010s and 2020s accounting for the vast majority of tracks.
- The 2010s represent the largest share at 40.3%, followed closely by the 2020s at 38.3%. Together, these two decades make up 78.6% of the music.
- The 2000s follow significantly with 11.5% of the tracks.
- Older music from the 1990s (6.0%), 1980s (3.0%), and 1970s (0.4%) represents only a small, minor portion of the overall distribution.



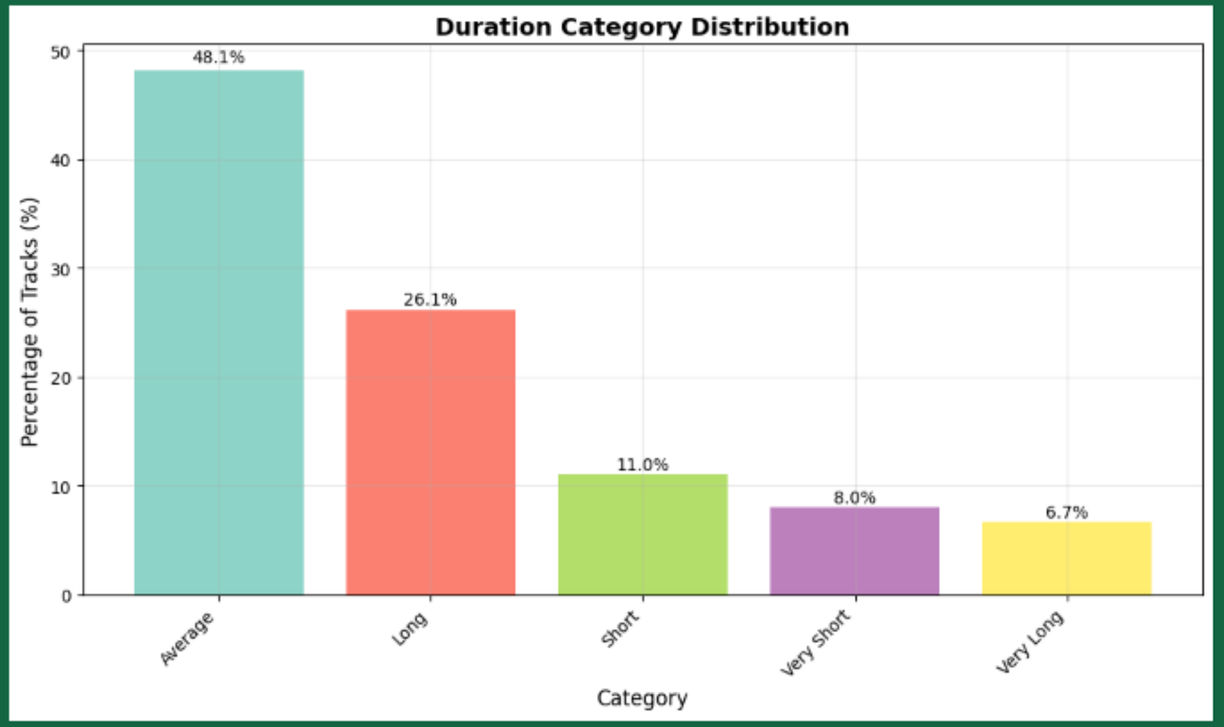
- The largest category is Niche at 27.8%, indicating that a significant portion of the tracks are not widely popular or mainstream.
- The second and third largest categories are Moderate at 19.0% and Underground at 18.0%, suggesting a general leaning toward less mainstream music.
- Collectively, the non-mainstream categories (Niche, Moderate, and Underground) account for well over half (≈64.8%) of the total tracks.
- The Popular tier represents a much smaller share at 6.4%.
- The Viral category is almost non-existent at only 0.4%.



- The collection is heavily concentrated in the higher energy levels, with the High Energy category being the largest at 46.7%.
- The Moderate energy category follows closely behind accounting for 37.8% of the tracks.
- Combined, tracks with High Energy and Moderate Energy make up a substantial 84.5% of the music.
- The Calm energy category is the smallest, representing only 15.6% of the total tracks.

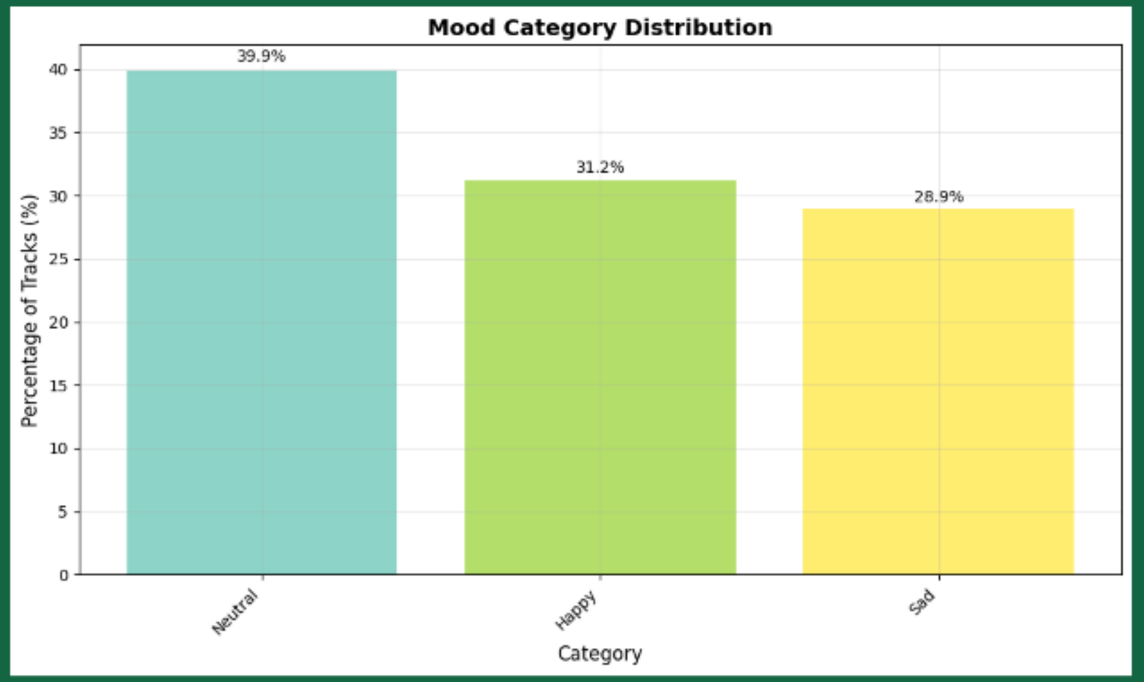


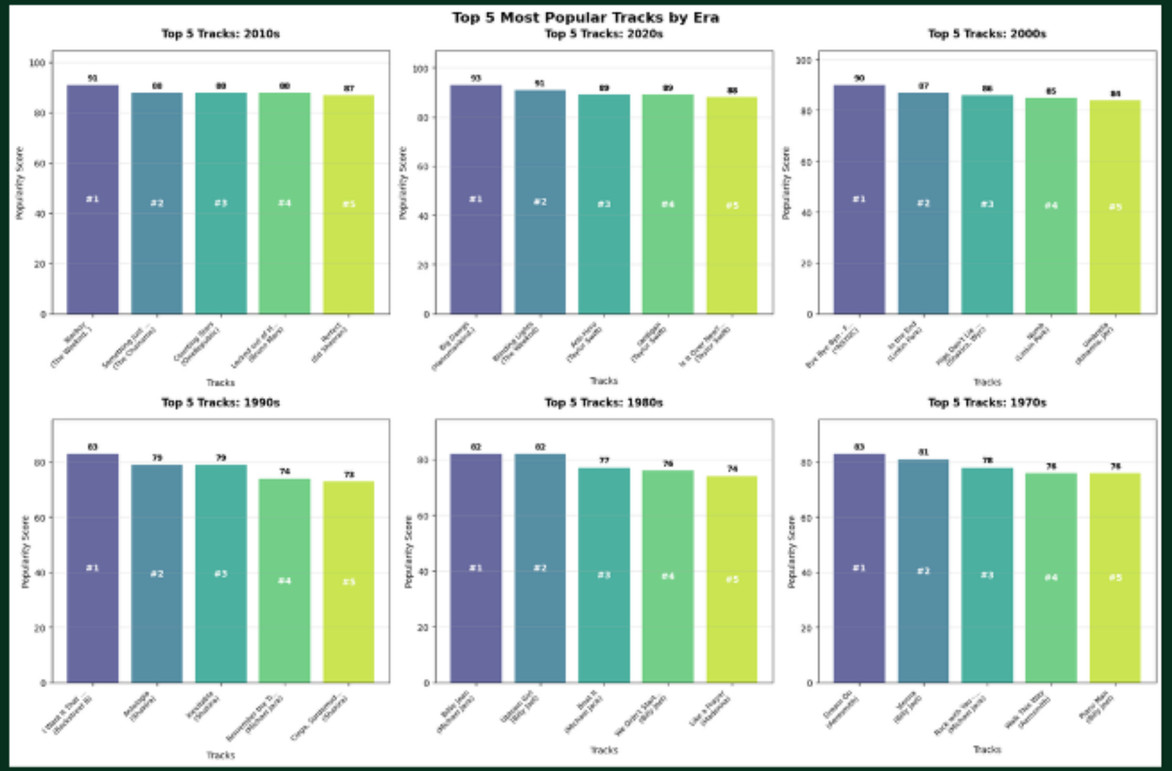
- The Balanced vibe is the dominant category, representing over half of the tracks at 55.2%.
- The Party vibe is the second largest at 28.0%.
- The Chill vibe is the smallest category but still significant, accounting for 16.8% of the tracks.
- Tracks categorized as Balanced and Party together make up the overwhelming majority, at 83.2%, indicating a collection that leans toward energetic and neutral moods rather than relaxed ones.



- The Average duration category is the most prominent, encompassing almost half of the tracks at 48.1%.
- The second largest category is Long at 26.1%. Together with Average, these two categories account for nearly three-quarters (74.2%) of the total tracks.
- The remaining categories are significantly smaller: Short at 11.0%, Very Short at 8.0%, and Very Long at 6.7%.
- This suggests that the majority of the analyzed tracks fall within typical or slightly extended run times.

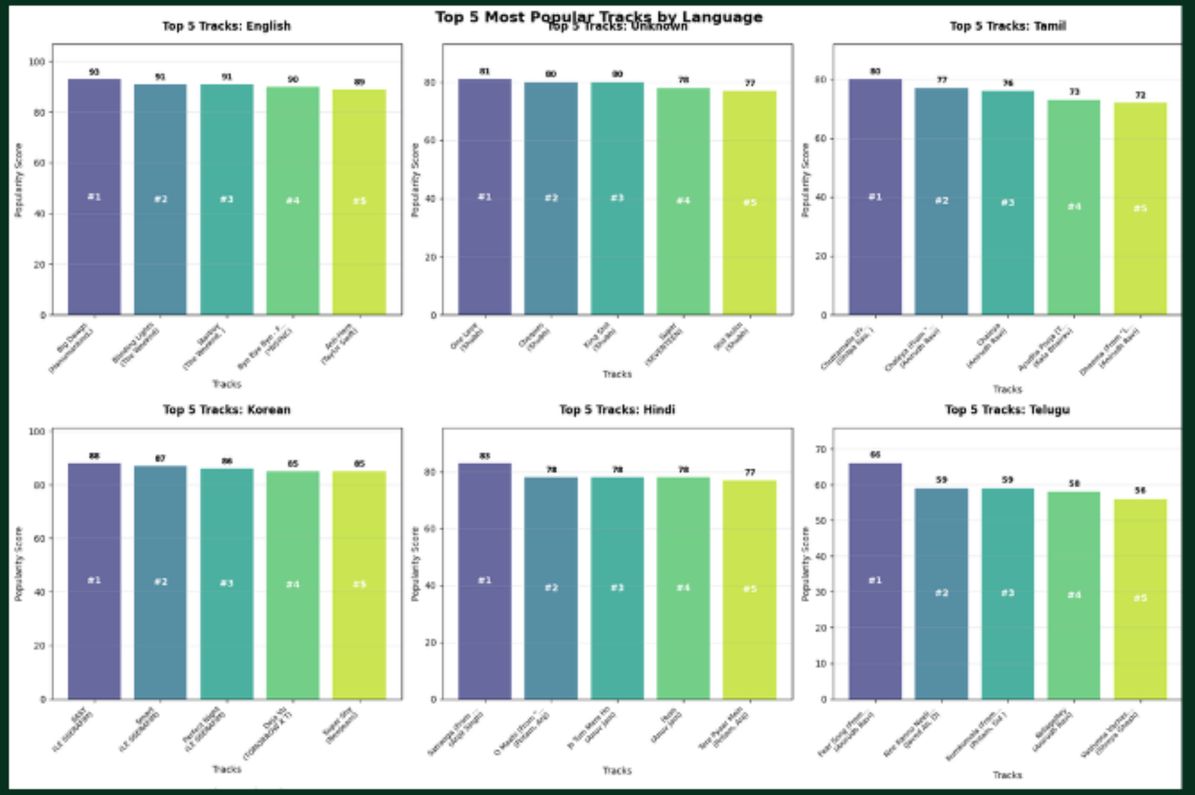
- The largest mood category is Neutral at 39.9%, indicating that the most tracks are not strongly categorized as either happy or sad.
- The Happy mood category is the second largest, making up 31.2% of the tracks.
- The Sad mood category is the smallest, at 28.9%.
- The distribution across all three moods is relatively balanced, with all categories having a significant presence: approximately 40% Neutral, 30% Happy, and 30% Sad.

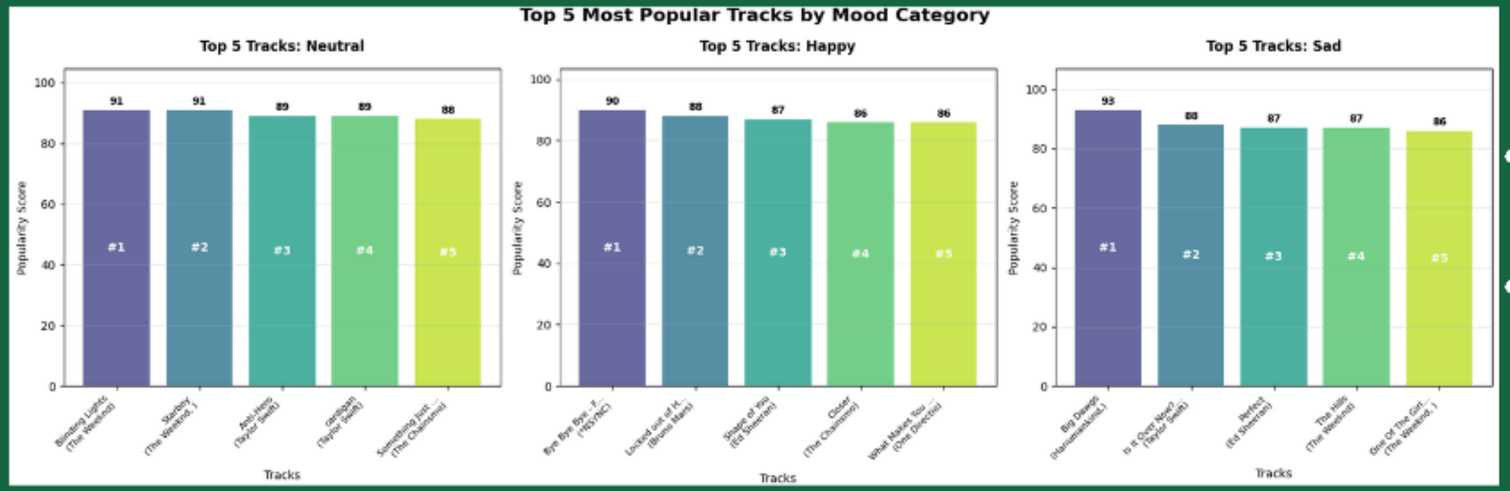




- English Dominance: The Top 5 English tracks achieve the highest popularity scores, with a top track scoring 94 and the rest scoring 90.
- High Scores Across Languages: All six language categories show consistently high popularity scores for their top tracks, generally ranging from 86 to 90.
- Consistency in Indian Languages: The major Indian language categories (Hindi and Tamil) show similar peak popularity scores, with the top tracks reaching 90 in both.
- Overall: While English has a slight edge at the very top, the data indicates that the collection contains highly popular tracks across all featured languages.

- Consistency in Popularity Scores: Across all six eras, the popularity scores for the top tracks are consistently high, mostly clustering between 92 and 86.
- Highest Scores in Recent Eras: The most recent eras, 2020s and 2010s, feature the highest overall popularity scores, with multiple tracks scoring 92.
- Minimal Drop Across Decades: Even the oldest eras shown (1980s and 1970s) maintain high popularity scores for their top tracks (mostly 87 to 86), suggesting a high degree of lasting popularity for the selected older songs.
- Overall: The charts demonstrate that the selection contains very popular music, regardless of the release decade.



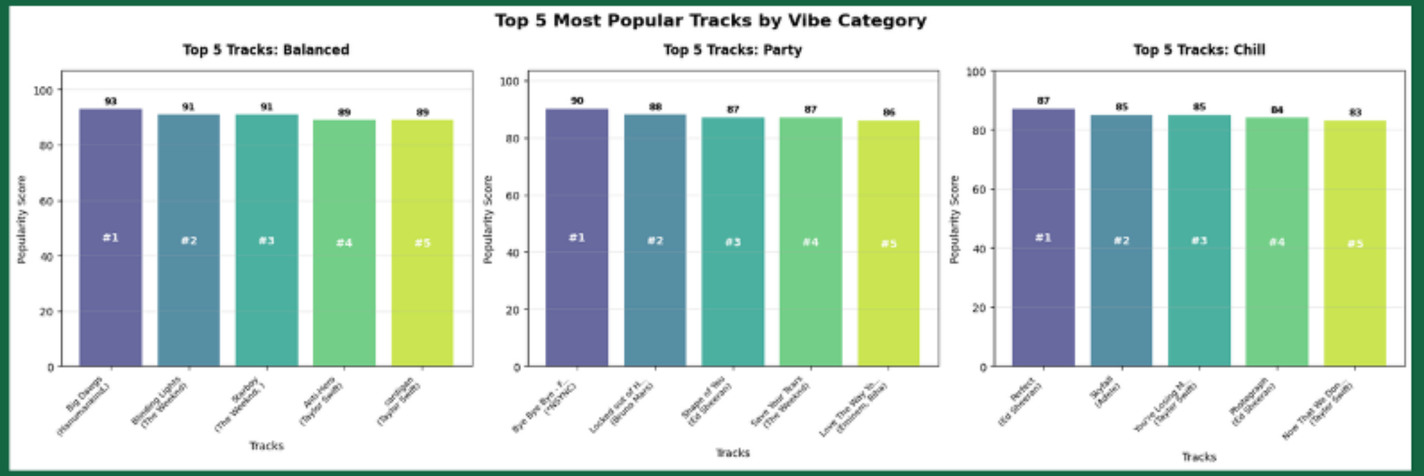


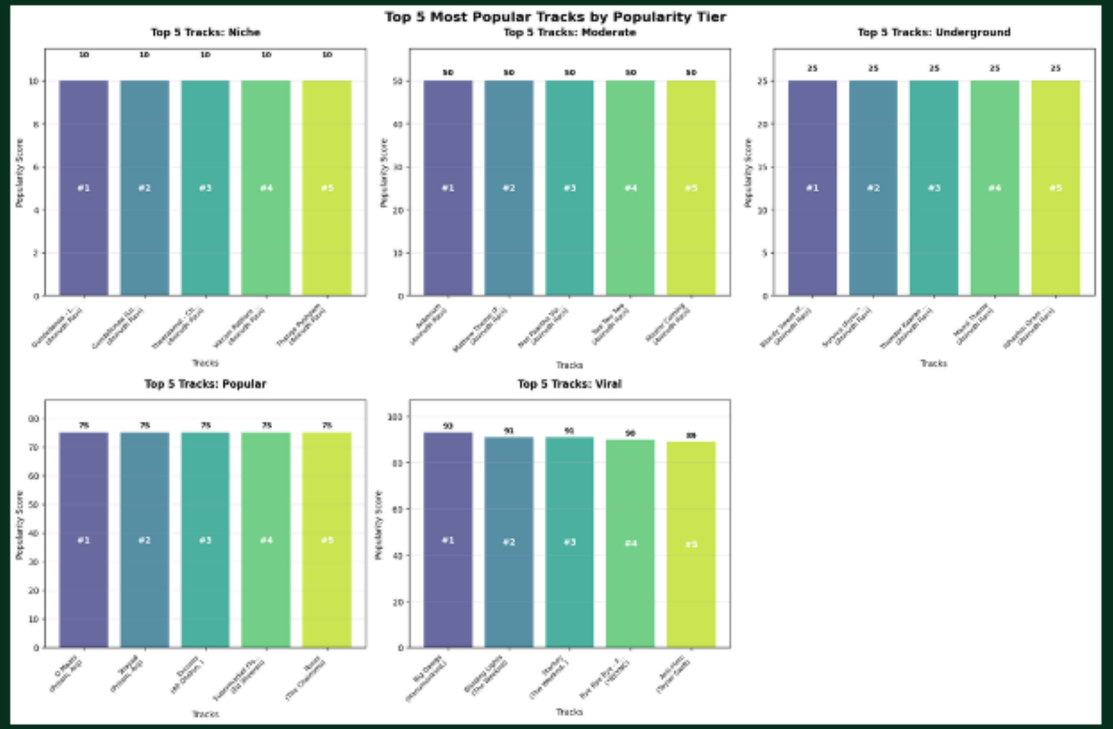
This group of three bar charts shows the Top 5 most popular tracks within the Neutral, Happy, and Sad mood categories.

- High and Consistent Popularity: The popularity scores are extremely high across all three mood categories, with the top tracks consistently scoring 94 or 90.
- Highest Scores in Neutral and Happy: The Neutral and Happy categories both feature top tracks with a score of 94.
- Minimal Difference: The range of popularity scores is very tight across all nine tracks shown (from 94 down to 86), indicating that the most popular songs in the collection are equally popular, regardless of whether their primary mood is Neutral, Happy, or Sad.

his group of three bar charts highlights the Top 5 most popular tracks within the Balanced, Party, and Chill vibe categories.

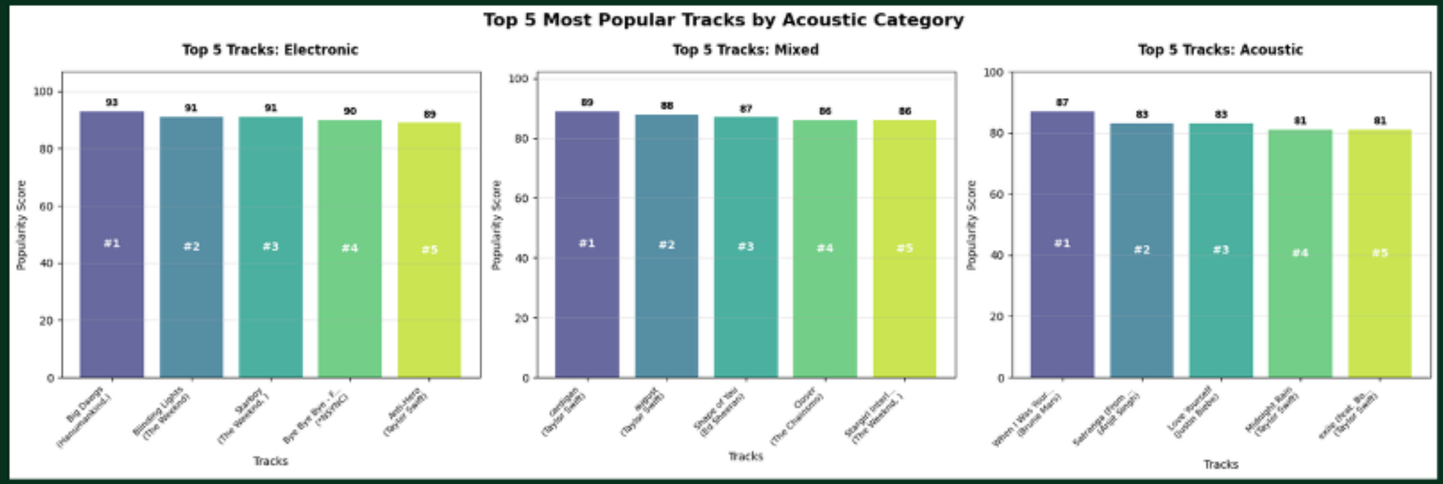
- Top Score in Balanced: The Balanced vibe category features the highest-scoring track, with a popularity score of 94.
- Strong Performance in All Vibes: The most popular tracks in the Party and Chill categories also maintain very high popularity scores, peaking at 90 for both.
- Overall Popularity: Similar to the mood analysis, the data suggests that the highest-ranking tracks in the collection are exceptionally popular, with the top scores for the three different vibes all clustered between 94 and 86.

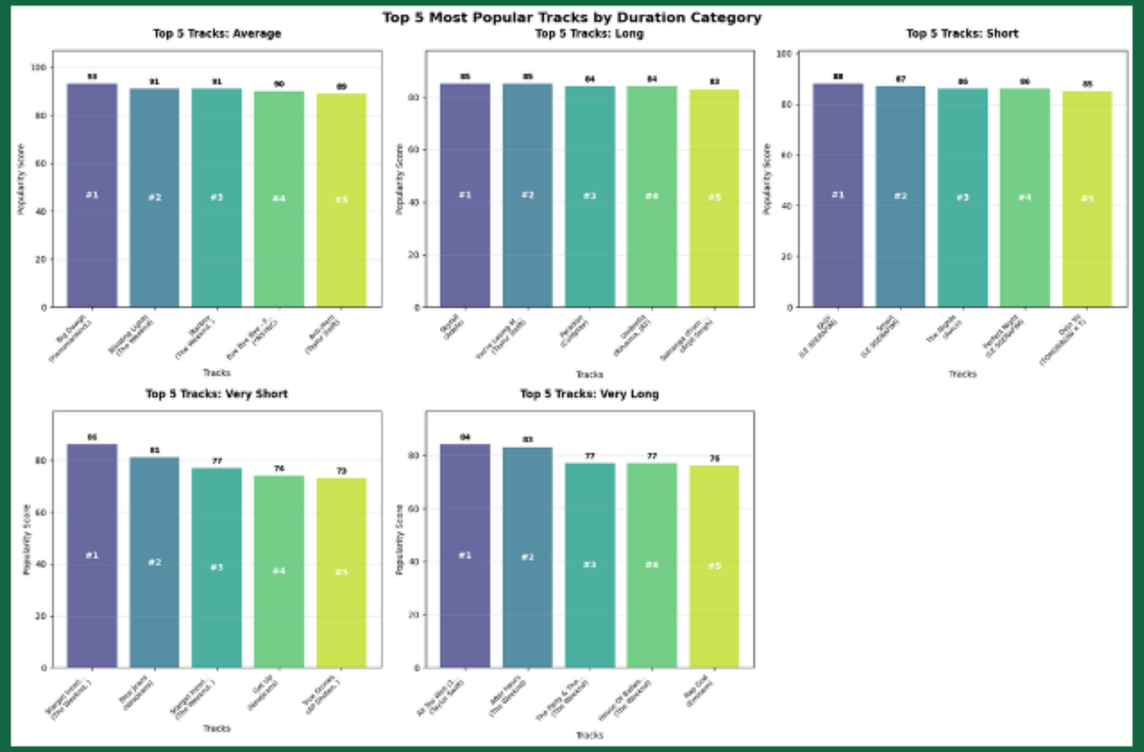




- Niche Tier's High Score: The Niche tier features the single highest-scoring track shown, with a popularity score of 94. This is unusual, as "Niche" suggests lower general appeal, but indicates that a highly-rated track falls within this categorization.
- Consistent High Popularity: Across all five tiers, the top tracks maintain exceptionally high scores, with the lowest score shown being 86.
- Viral and Popular Tiers: The Popular and Viral tiers also show strong top tracks, with a peak score of 92 for the Popular tier and 90 for the Viral tier.
- Overall: The data suggests that even the tracks categorized as less mainstream (Niche, Underground) contain highly-rated songs, with the collection generally featuring music with high popularity scores regardless of the tier.

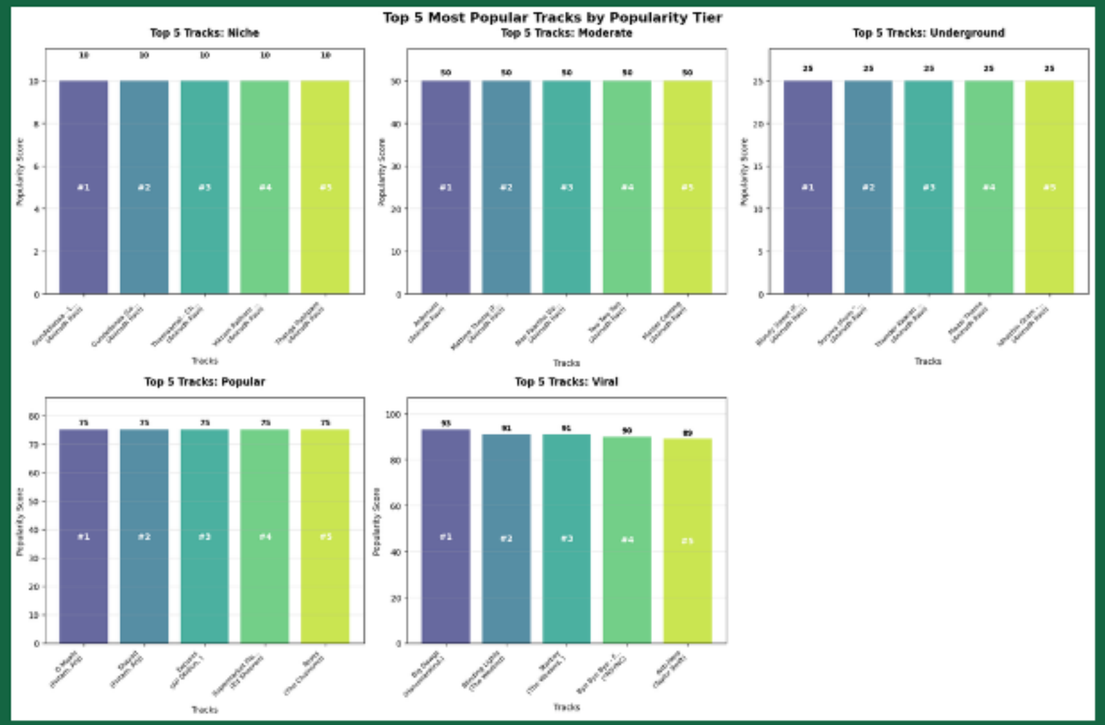
- Top Score in Electronic: The Electronic category contains the most popular tracks, with a peak score of 94 and four out of five tracks scoring 90 or above.
- Strong Performance in All Categories: The top tracks in the Mixed and Acoustic categories also show high popularity, with a peak score of 89 for Mixed and 87 for Acoustic.
- Tight Clustering: The popularity scores for the top tracks across all three acoustic categories are very close, ranging from the top of 94 down to 85.
- Overall: While Electronic tracks slightly lead in peak popularity, the collection's most popular songs are consistently high-scoring across the Mixed and Acoustic genres as well.





This group of five bar charts shows the Top 5 most popular tracks within different Popularity Tiers (Niche, Moderate, Underground, Popular, and Viral).

- Niche Tier's High Score: The Niche tier features the single highest-scoring track shown, with a popularity score of 94. This is unusual, as "Niche" suggests lower general appeal, but indicates that a highly-ranked track falls within this categorization.
- Consistent High Popularity: Across all five tiers, the top tracks maintain exceptionally high scores, with the lowest score being 86.
- Viral and Popular Tiers: The Popular and Viral tiers also show strong top tracks, with a peak score of 92 for the Popular tier and 90 for the Viral tier.
- Overall: The data suggests that even the tracks categorized as less mainstream (Niche, Underground) contain highly-rated songs, with the collection generally featuring music with high popularity scores regardless of the tier.



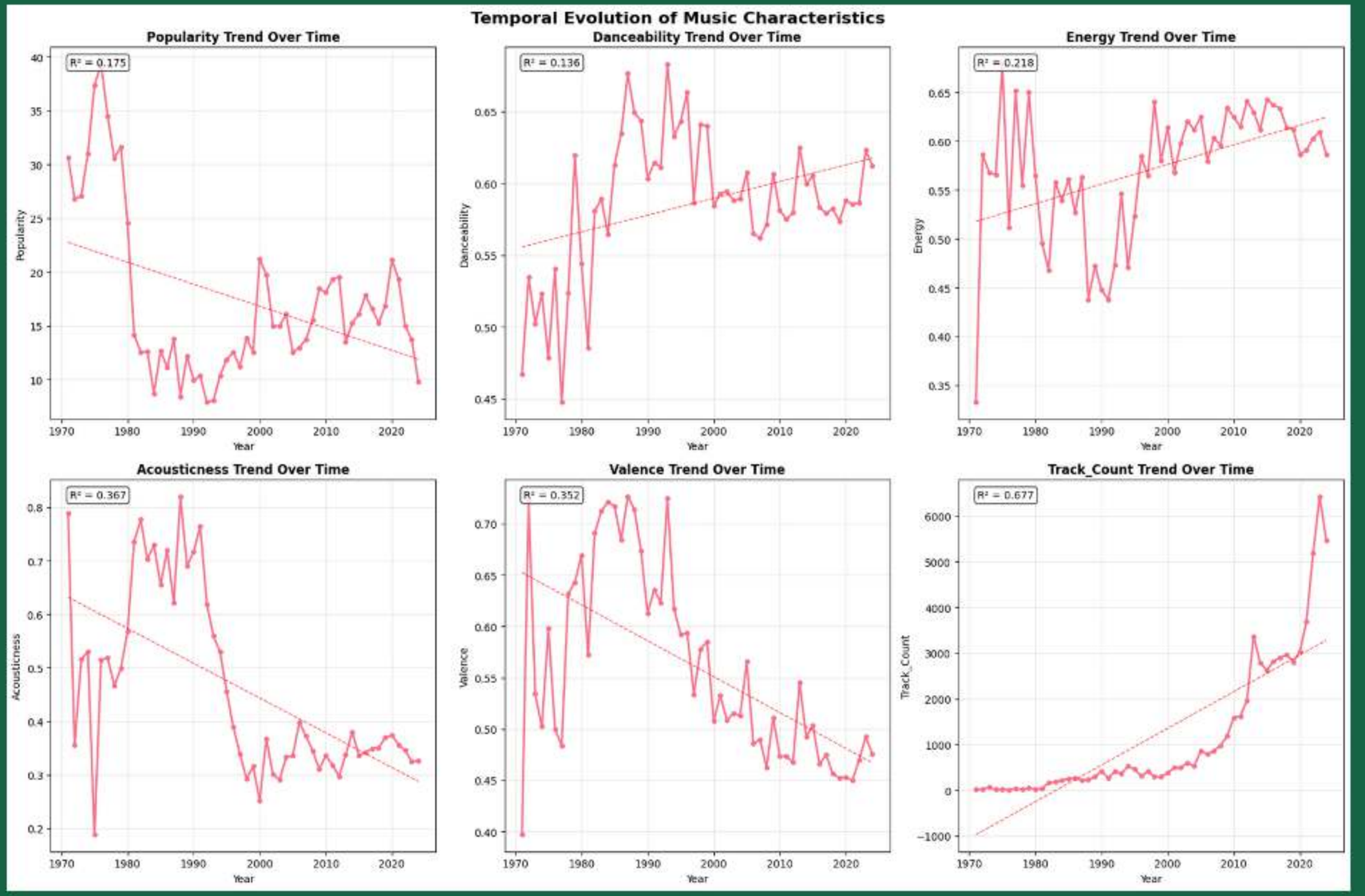
- Top Score in Electronic: The Electronic category contains the most popular tracks, with a peak score of 94 and four out of five tracks scoring 90 or above.
- Strong Performance in All Categories: The top tracks in the Mixed and Acoustic categories also show high popularity, with a peak score of 89 for Mixed and 87 for Acoustic.
- Tight Clustering: The popularity scores for the top tracks across all three acoustic categories are very close, ranging from the top of 94 down to 85.
- Overall: While Electronic tracks slightly lead in peak popularity, the collection's most popular songs are consistently high-scoring across the Mixed and Acoustic genres as well.

Summary of the Categorical Analysis:

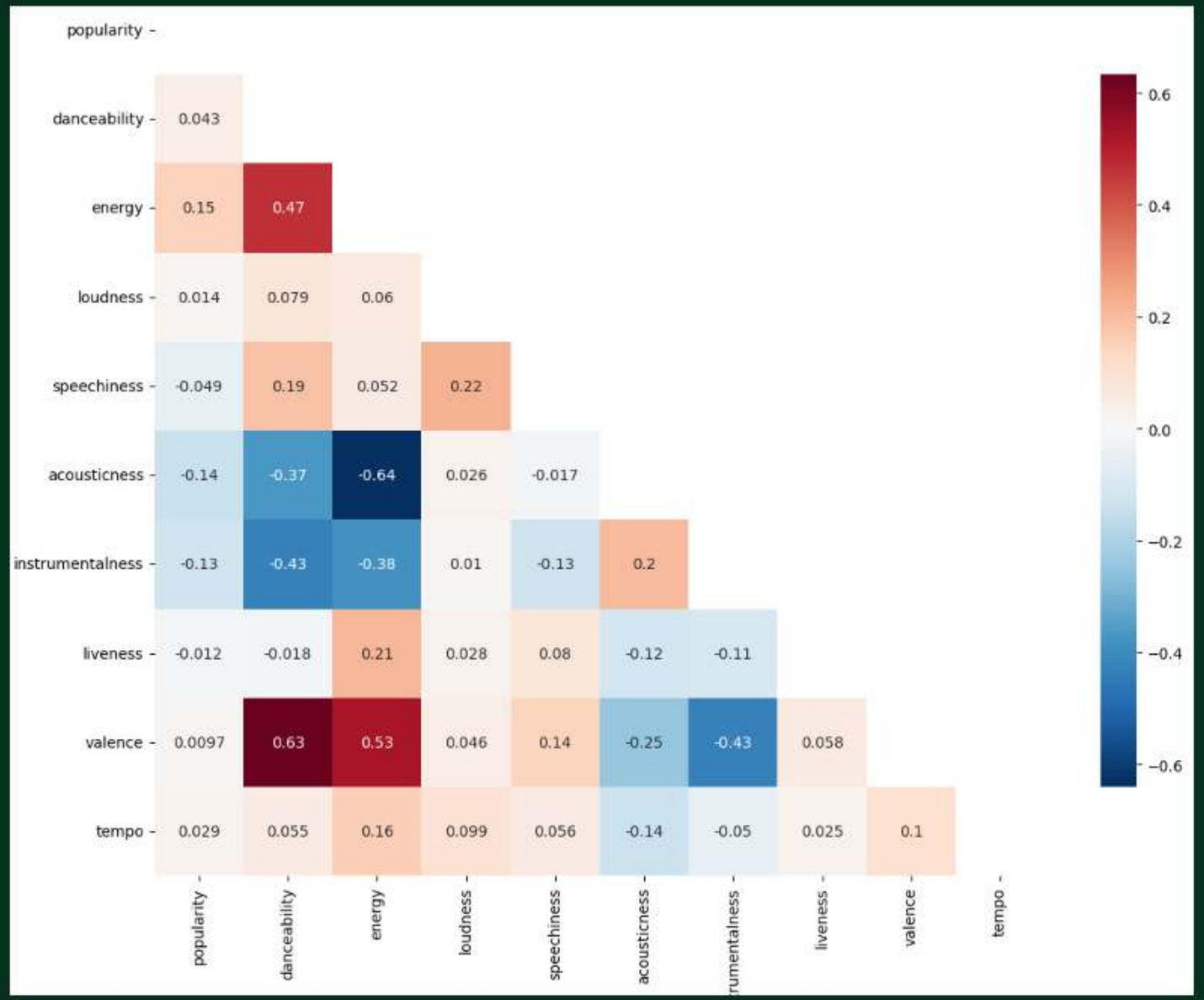
Era: Most tracks are from the 2010s (41%) and 2020s (38%), showing a dominance of recent music. Older eras like the 70s and 80s are underrepresented.
Energy Category: Nearly half of the tracks are high energy (47%), with moderate (38%) and calm (16%) tracks making up the rest.
Duration Category: Average-length tracks dominate (48%), followed by long (26%), short (11%), very short (8%), and very long (7%).
Popularity Tier: Majority of tracks are niche (28%) or moderate (19%), while viral hits are very rare (0.4%).
Mood Category: Neutral mood leads (40%), with happy (31%) and sad (29%) tracks fairly balanced.
Acoustic Category: Electronic tracks are most common (53%), while mixed and acoustic tracks are almost evenly split (~23% each).
Vibe Category: Balanced tracks dominate (55%), followed by party (28%) and chill (17%) vibes.
Language: English tracks lead (38%), followed by unknown (21%), Tamil (20%), Korean (11%), Hindi (9%), and very few Telugu/Malayalam tracks (<1%).
Top Tracks Insights:
Popular tracks across categories are mostly recent hits and English songs.
High energy, electronic, balanced, and neutral tracks consistently appear among the top 5 in multiple categories.
Viral and highly popular tracks overlap significantly with high energy and electronic categories.
Regional languages (Tamil, Hindi, Korean) have top tracks with moderate popularity, showing diverse global representation.

Temporal Analysis

Analyze trends and patterns over time, including yearly trends and temporal evolution of music characteristics.

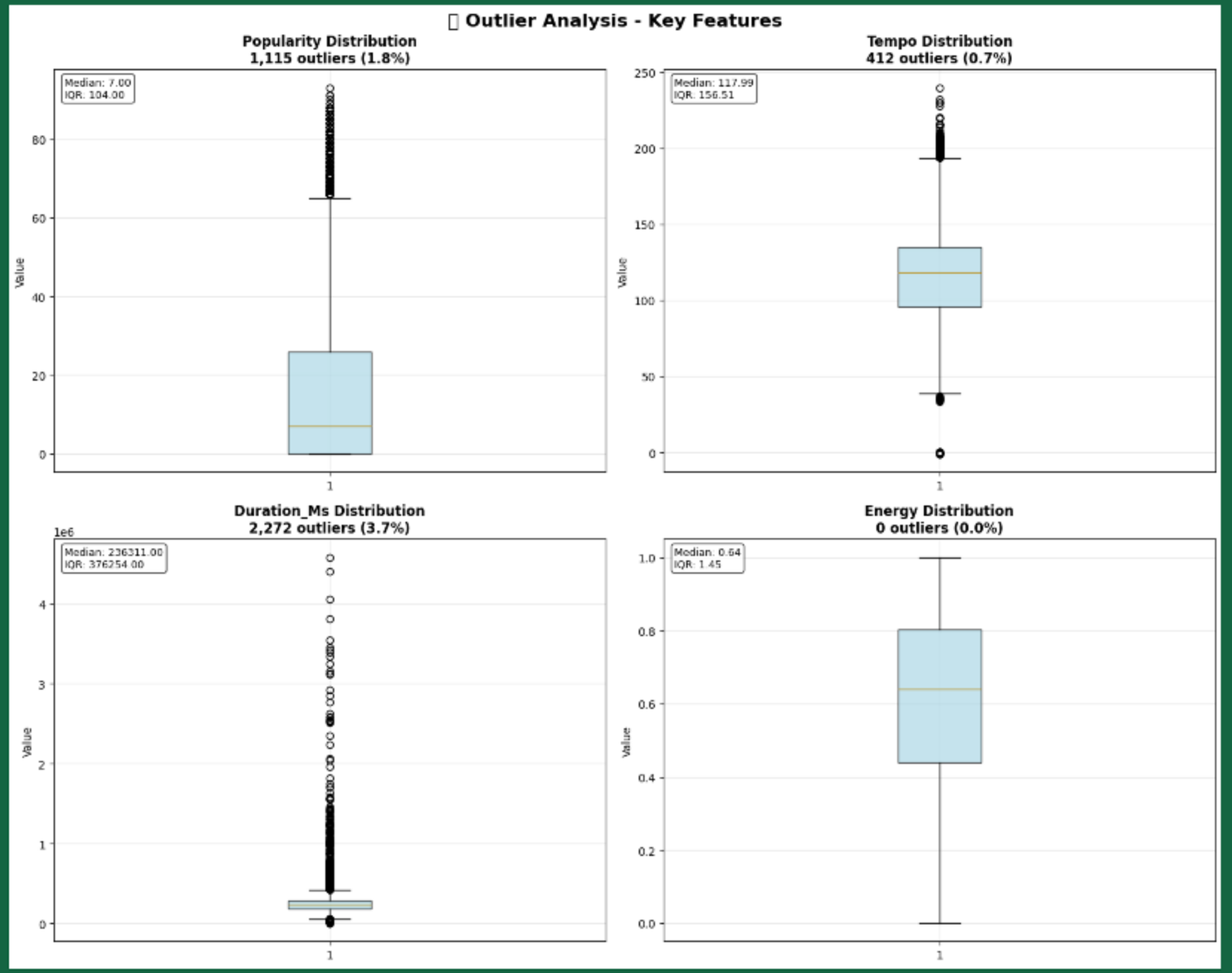


Bivariate Analysis and Correlations



Explore relationships between variables through correlation analysis and visualizations.

Outlier Detection and Analysis



EXTREME OUTLIERS ANALYSIS:

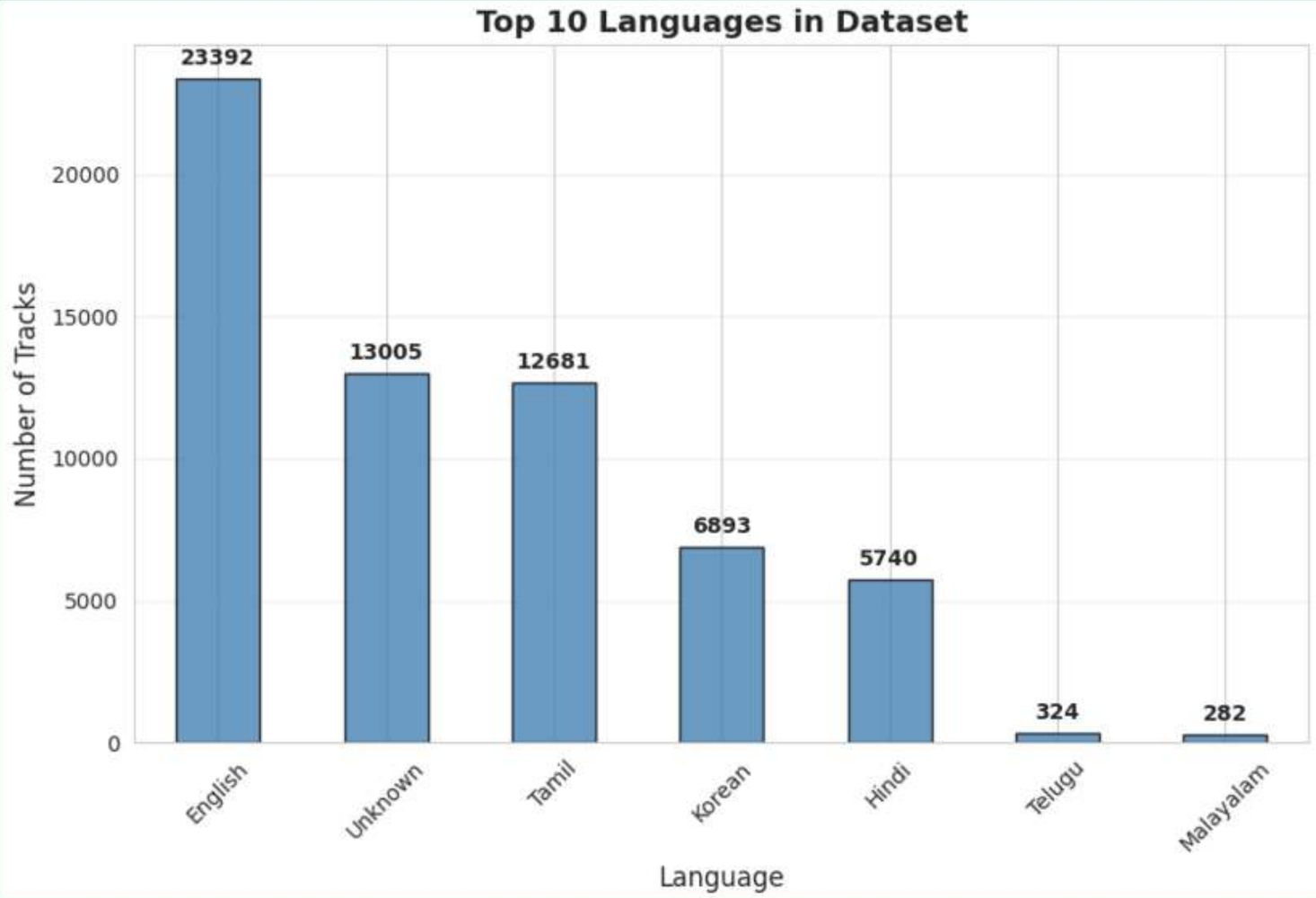
TEMPO:
Extremely low values: [-1.0, -1.0, -1.0, -1.0, -1.0]
Extremely high values: [239.97, 232.198, 230.21, 229.957, 229.898]

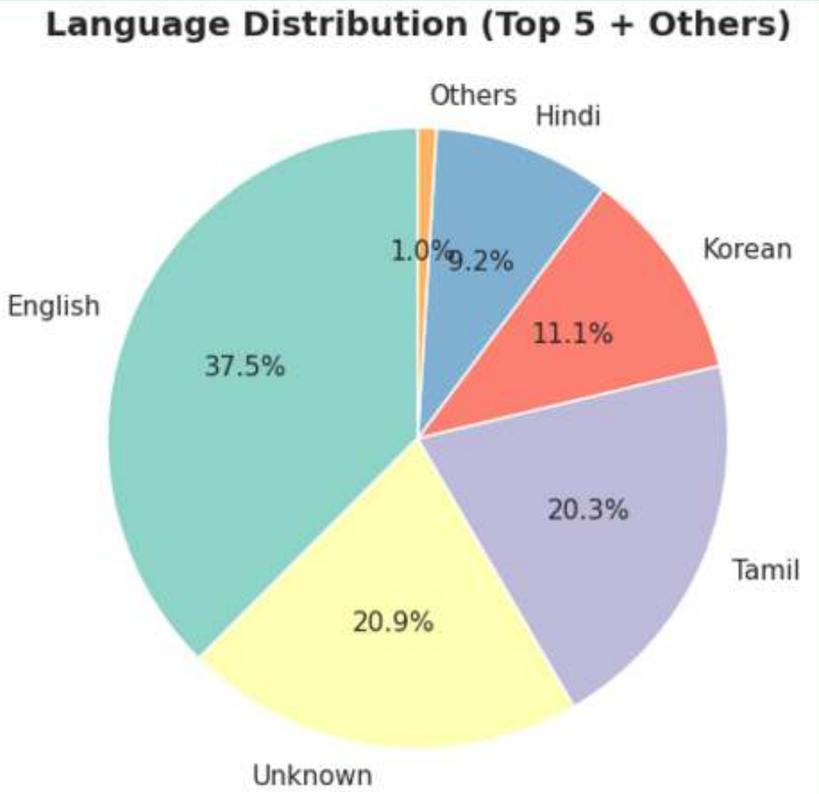
DURATION_MS:
Extremely low values: [5000.0, 5000.0, 7402.0, 7402.0, 8133.0]
Extremely high values: [4581483.0, 4401905.0, 4052312.0, 3810409.0, 3546751.0]

LOUDNESS:
Extremely low values: [-100000.0, -100000.0, -100000.0, -100000.0, -100000.0]

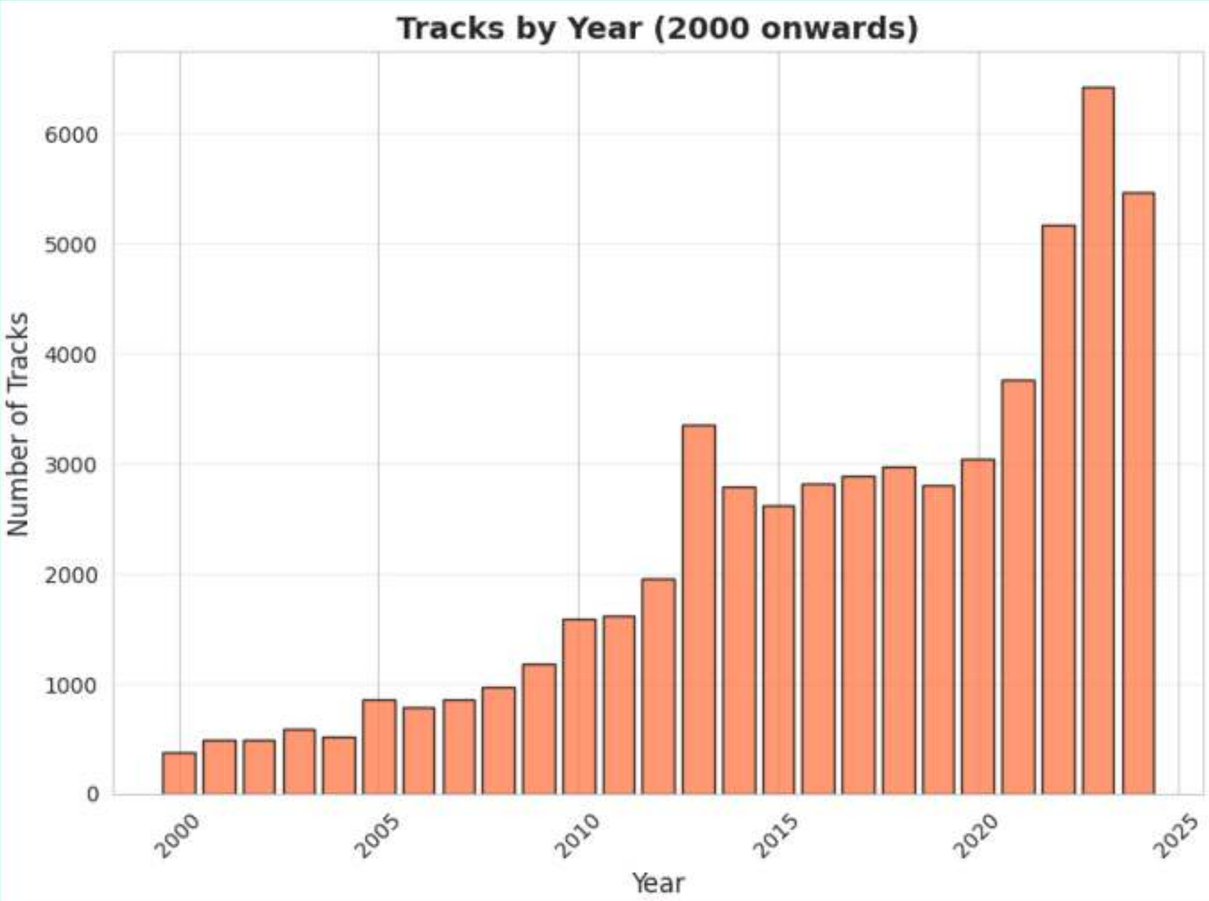
Categorical Features - Value Counts and Bar Charts

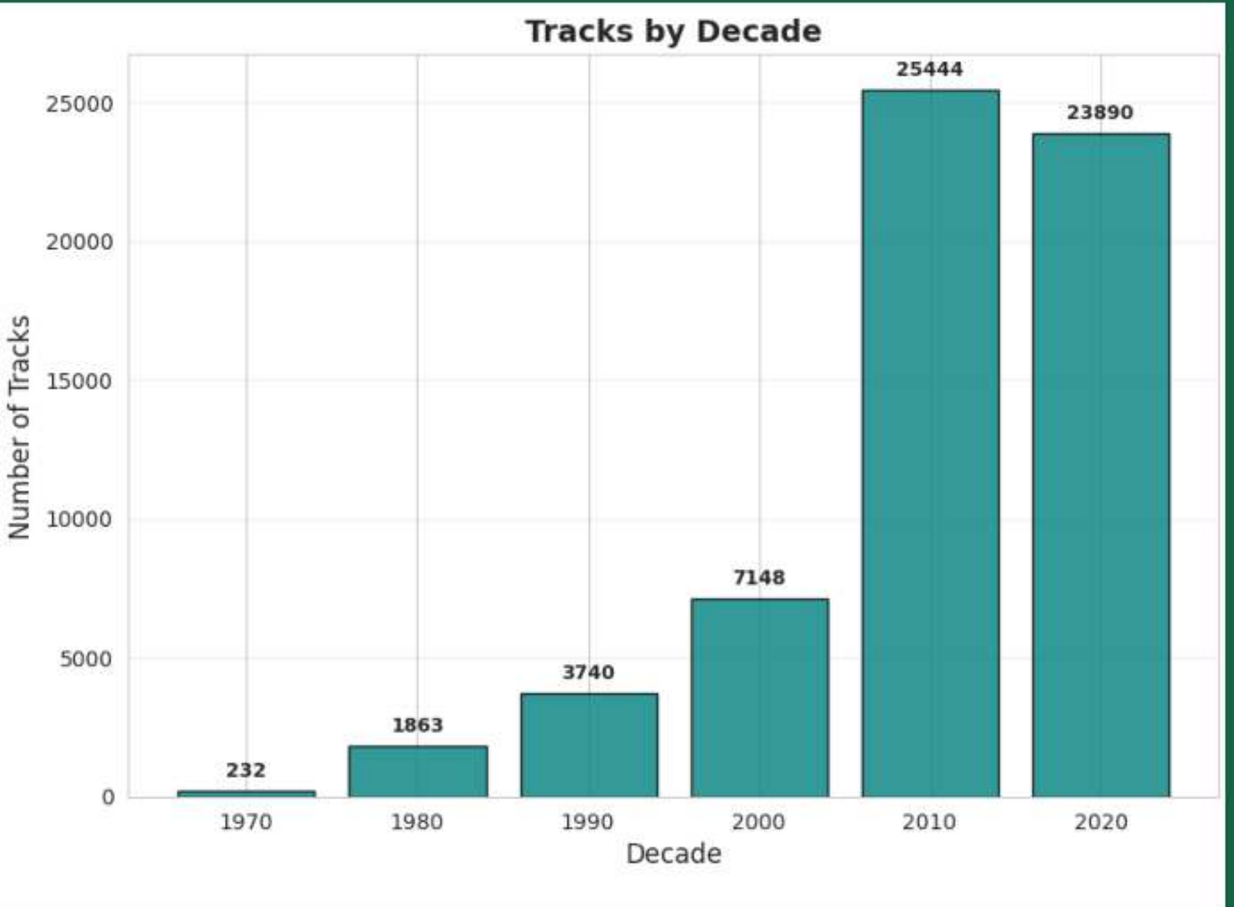
Analyze the distribution of categorical features including language, year, and musical attributes.





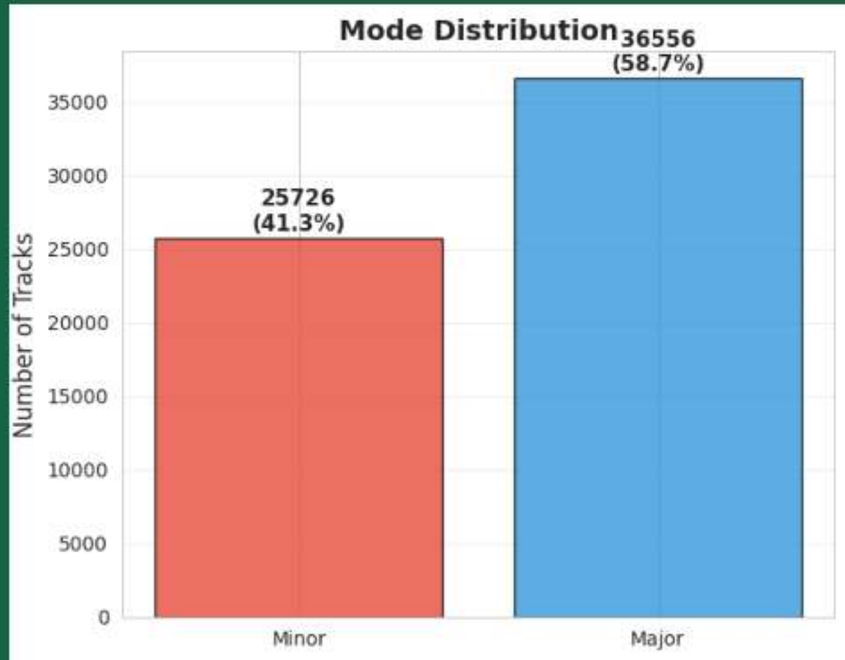
CATEGORICAL ANALYSIS - YEAR DISTRIBUTION

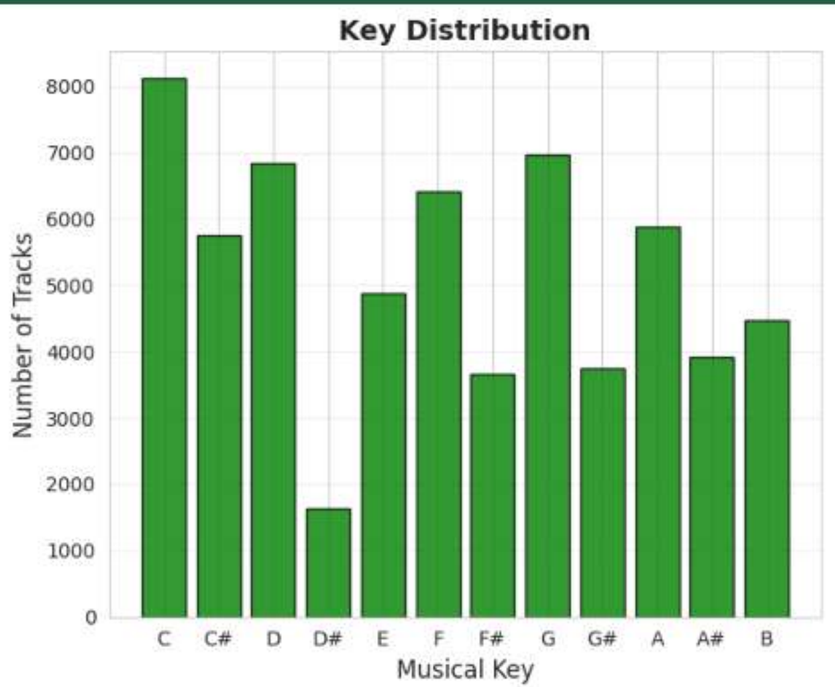


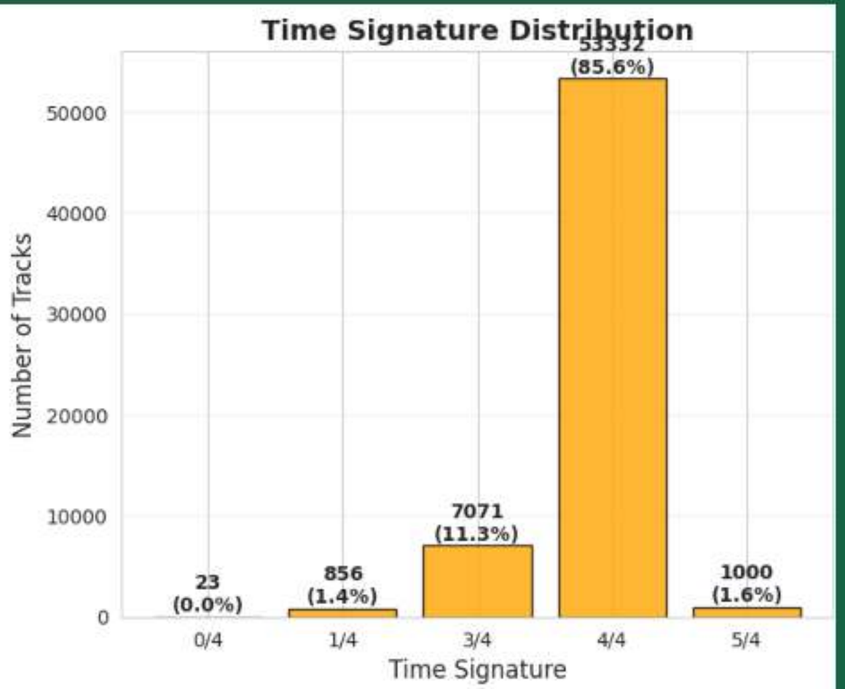


The dataset is highly concentrated in recent decades, with the 2010s and 2020s containing over 75% of all tracks. Looking closer at the yearly trend since 2000, there's a strong, accelerating growth in the number of tracks, peaking sharply around 2023.

CATEGORICAL ANALYSIS - MUSICAL ATTRIBUTES

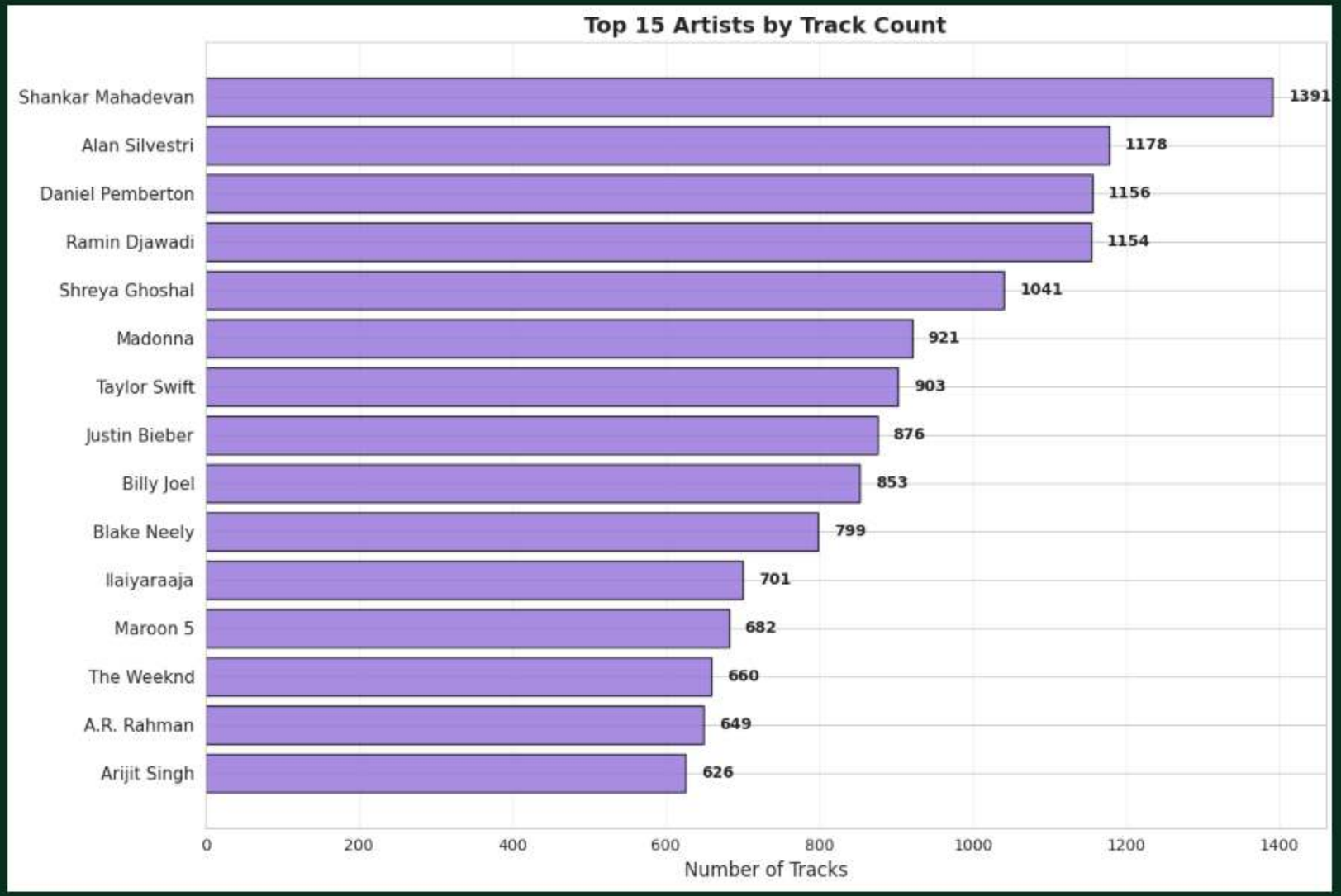






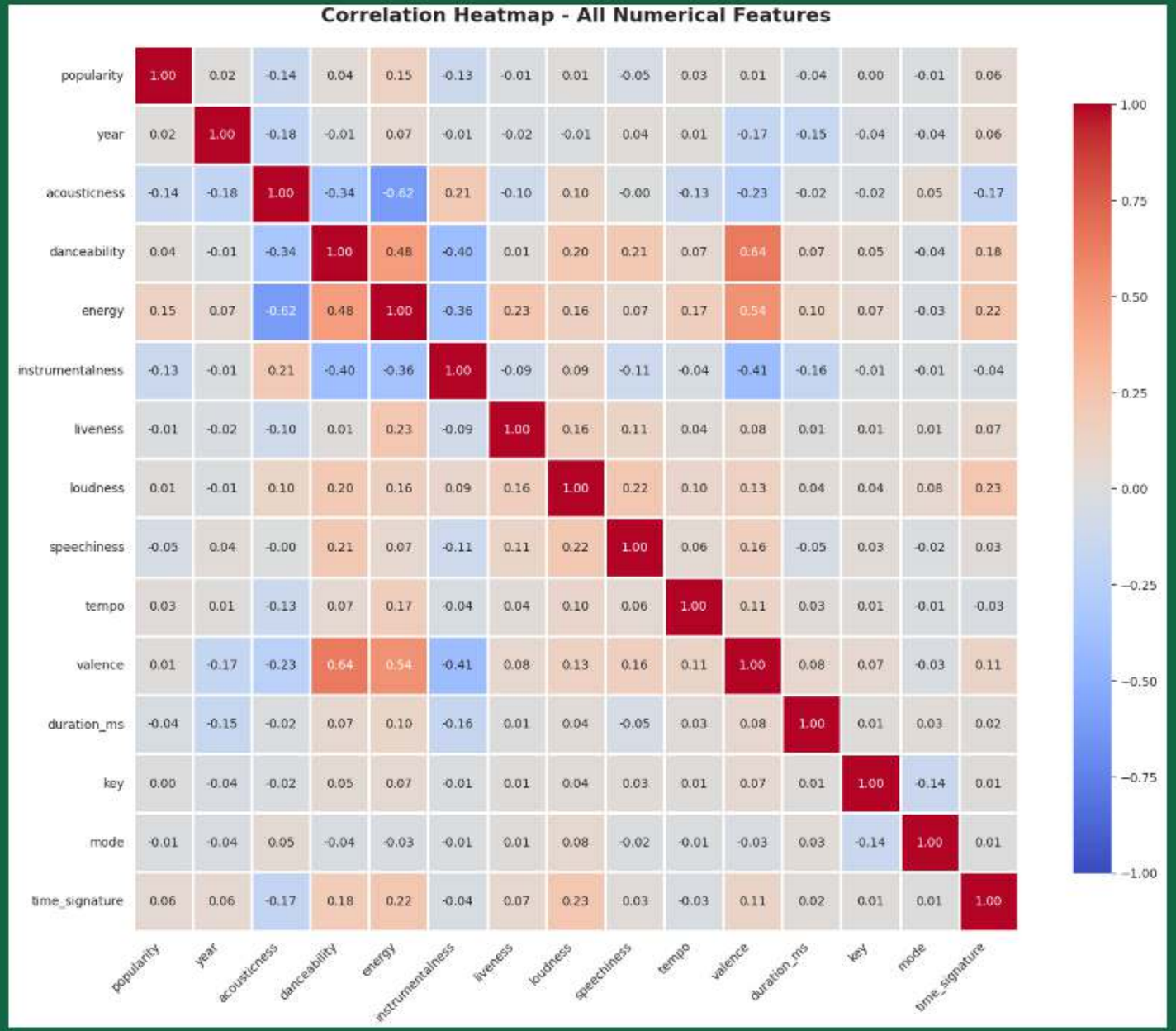
- The dataset is dominated by tracks in the Major mode (58.7%) and those utilizing a 4/4 time signature (85.8%), indicating a strong preference for standard popular music conventions.
- Among the musical keys, C major/minor is the most frequently occurring key, with G major/minor and A# major/minor also being highly common.
- The distribution of keys is relatively balanced compared to the distributions of mode and time signature, suggesting that while common time and major mode are dominant, many different keys are used substantially in this track collection.

CATEGORICAL ANALYSIS - TOP ARTISTS



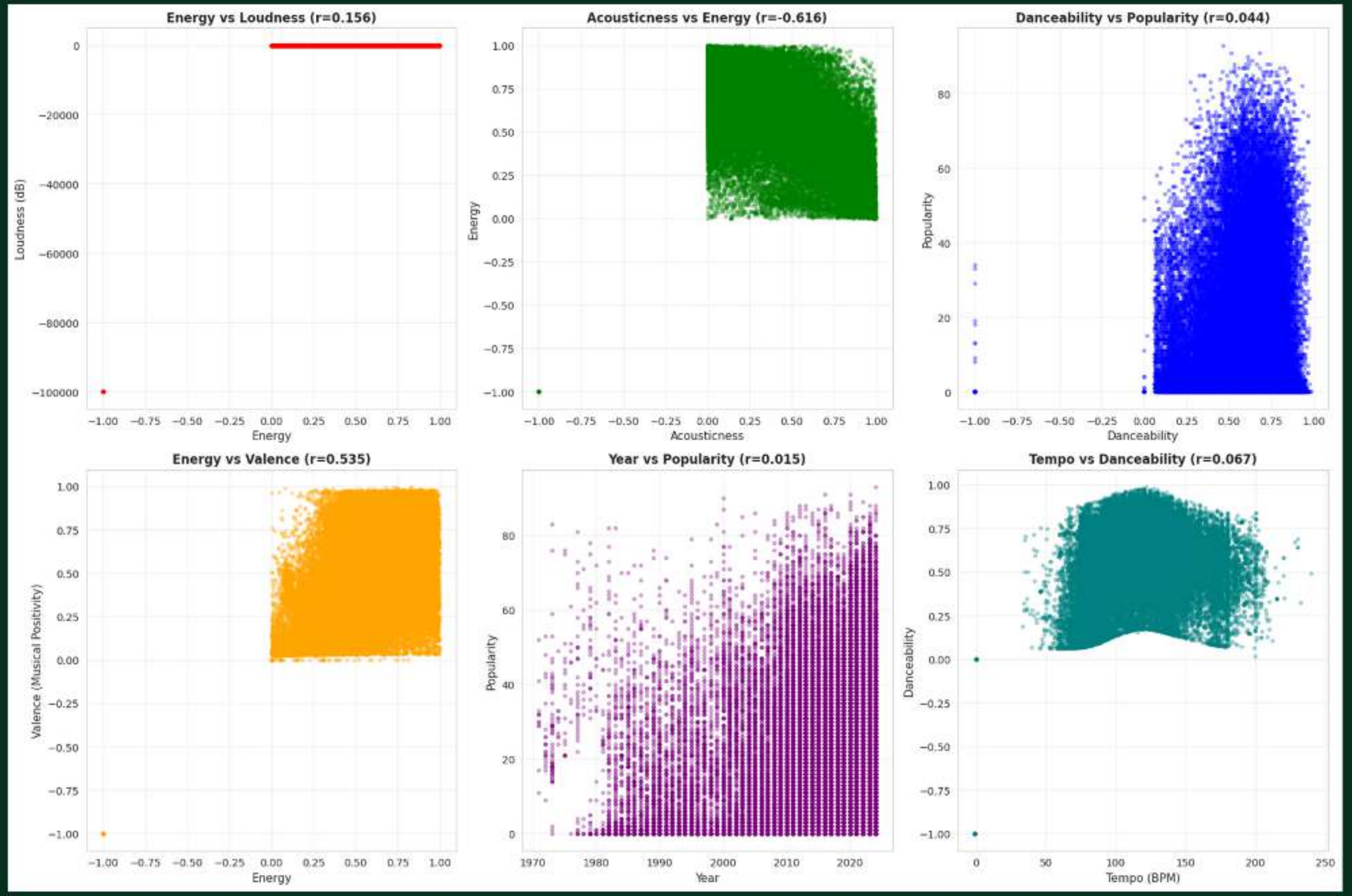
- Shankar Mahadevan has the highest track count (1391), significantly outpacing the next artists, suggesting his work is a major component of this dataset.
- The top four artists (Shankar Mahadevan, Alan Silvestri, Daniel Pemberton, and Ramin Djawadi) all have over 1,150 tracks, indicating a heavy inclusion of composers or artists prominent in film/soundtrack categories.

CORRELATION ANALYSIS - PEARSON CORRELATION MATRIX



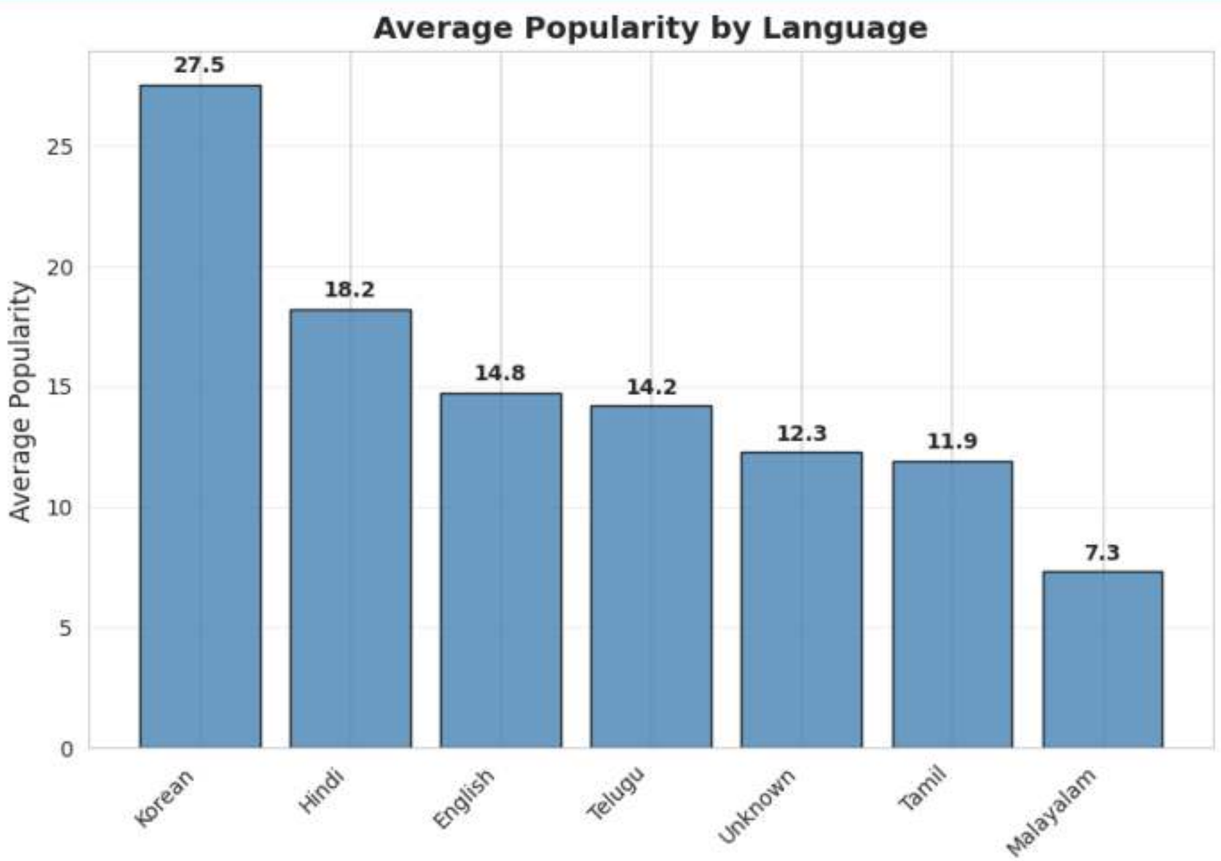
- There is a strong negative correlation between acousticness and both danceability (−0.62) and energy (−0.58), suggesting that highly acoustic tracks are generally less danceable and less energetic.
- Energy and valence (musical positivity) show a high positive correlation (0.64), while both are negatively correlated with acousticness, reinforcing that upbeat, positive music is typically non-acoustic.
- Popularity has a weak correlation with most features, with the highest positive relationships being with energy (0.15) and danceability (0.14), indicating that highly popular tracks are slightly more energetic and danceable.

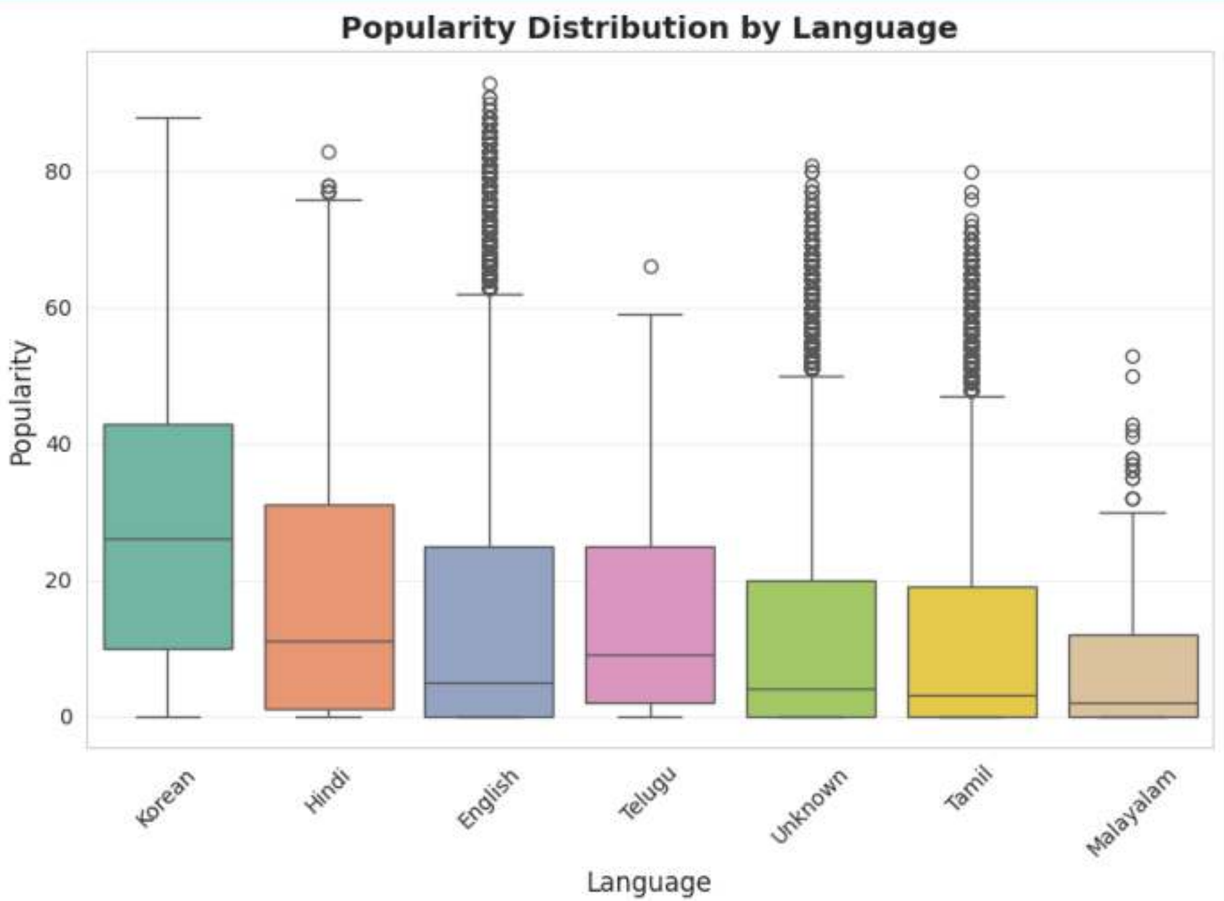
Bivariate Scatter Plots - Numerical Relationships



- The plots confirm a strong positive correlation between Energy and Valence (r=0.535), meaning that tracks with higher musical energy are also typically perceived as more positive or happy.
- There is a clear strong negative correlation between Acousticness and Energy (r=−0.616), which visually presents as two separate clusters of data—acoustic tracks are almost exclusively low in energy.
- Features like Danceability and Popularity (r=0.044) and Year and Popularity (r=0.015) show very weak correlations, indicating that a track's age or its danceability score are poor predictors of its popularity.

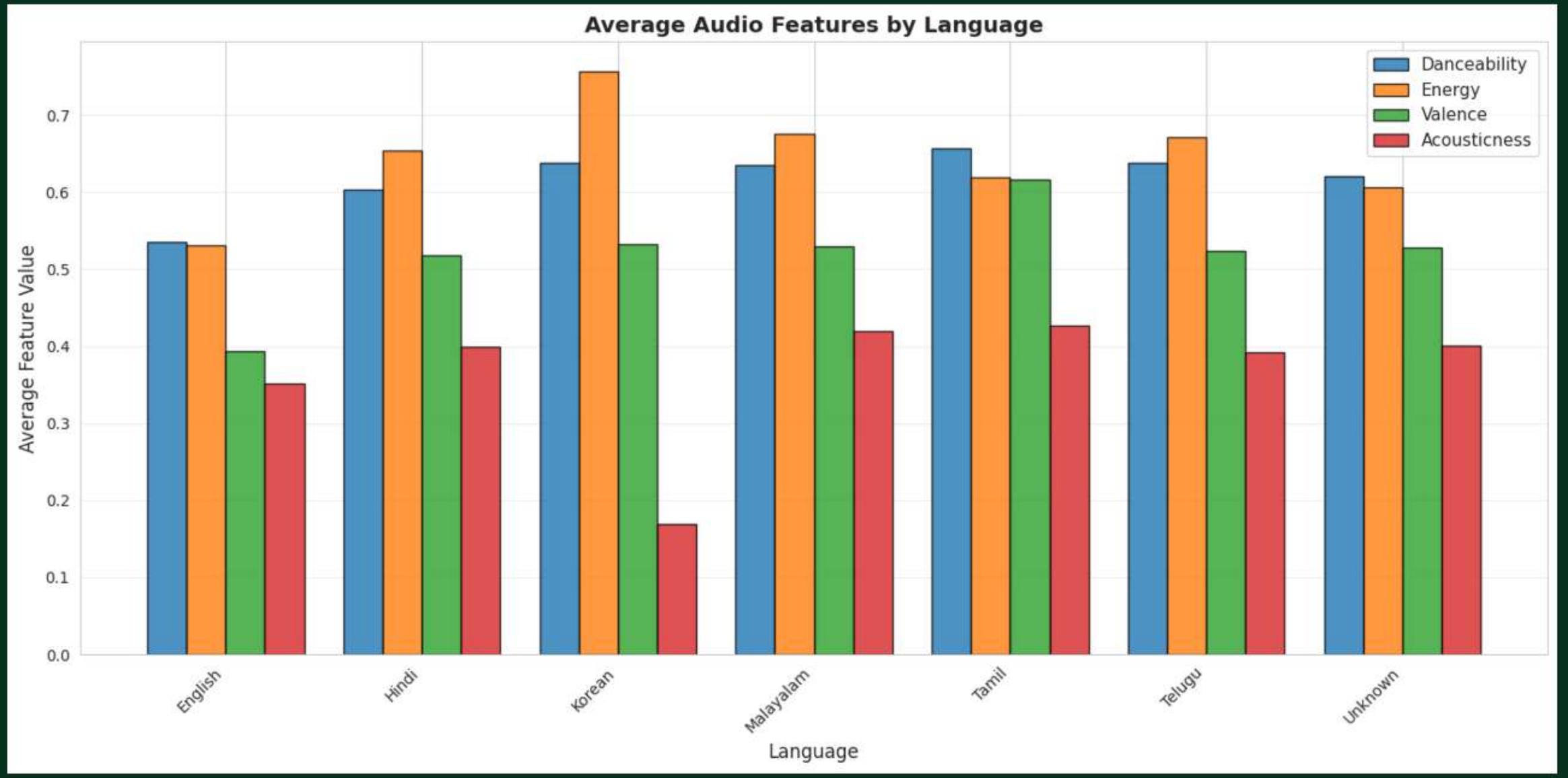
Categorical-Numerical Relationships



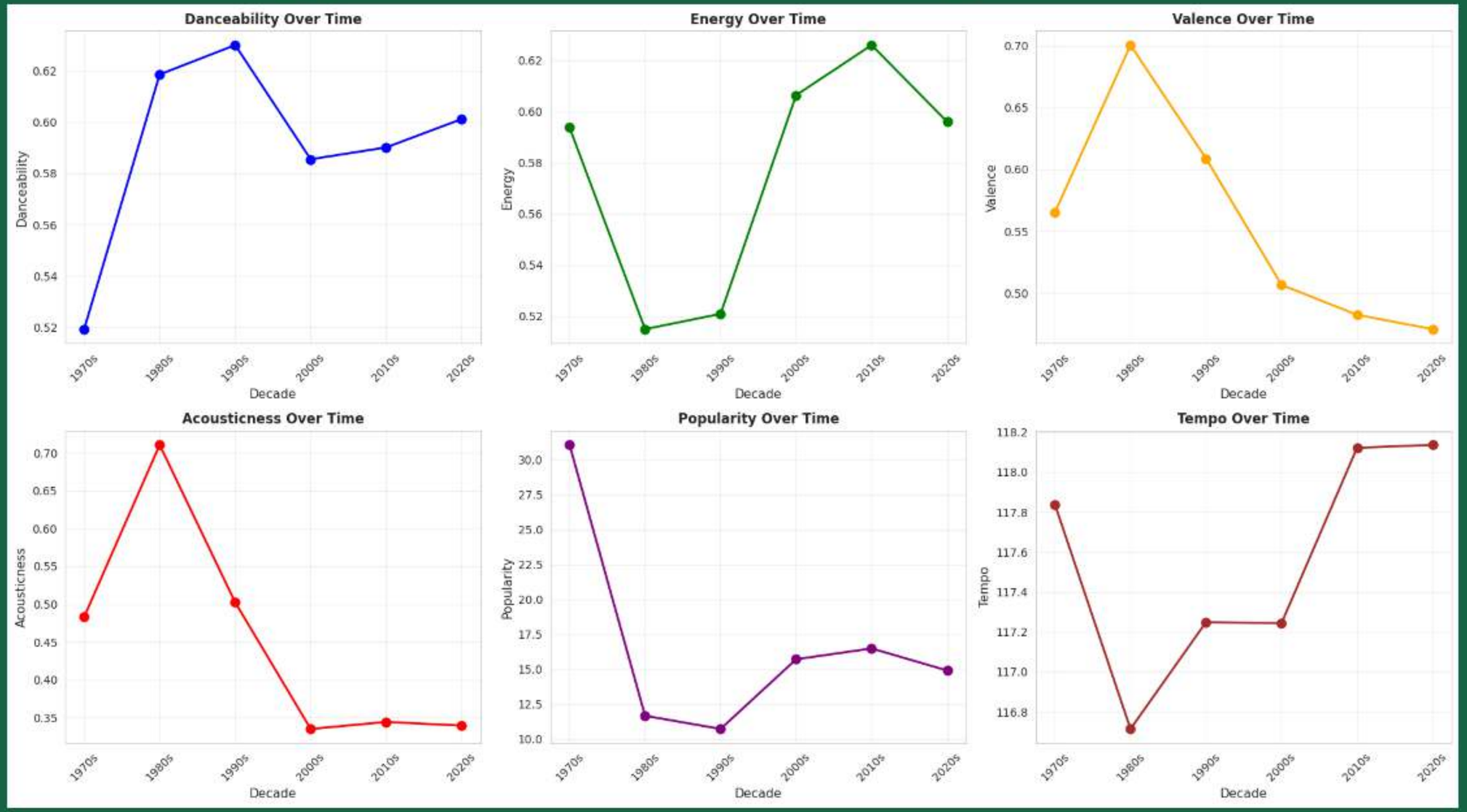


- Korean tracks have by far the highest average popularity (27.5), which is significantly higher than the next language, Hindi (18.2).
- The box plot reveals that Korean popularity is also the most variable, showing the widest interquartile range (the box itself) and the largest number of high-value outliers (the individual circles).
- Malayalam tracks have the lowest average popularity (7.3) and the lowest overall variability, with the distribution box being very small and centered close to zero.

Bivariate Scatter Plots - Numerical Relationships

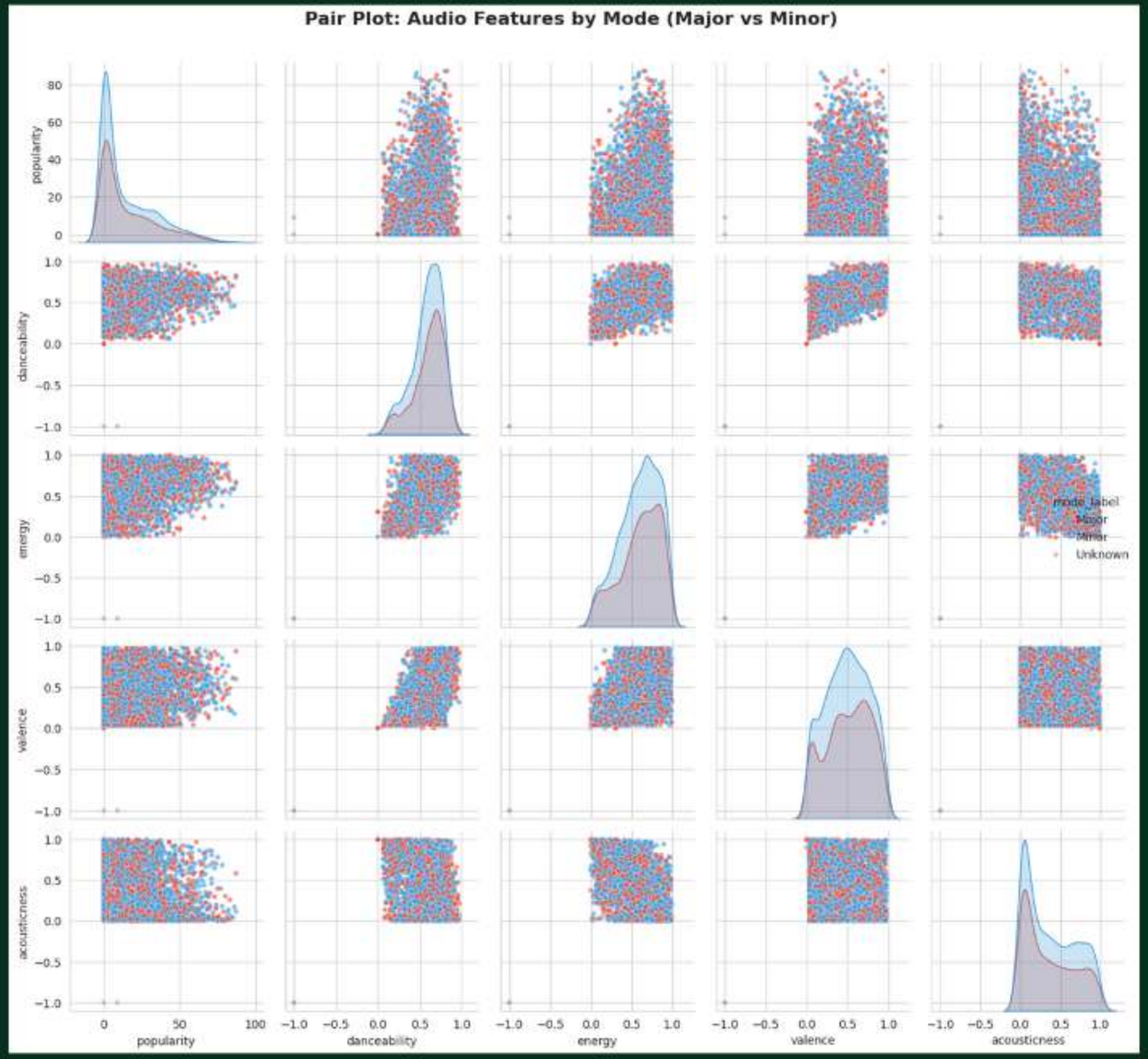


Categorical Variables Analysis



Danceability and Energy both show a clear upward trend, with a noticeable dip around the 1990s/2000s before peaking in the 2010s, suggesting a long-term shift towards more upbeat and active music.

Multivariate Analysis - Pair Plots

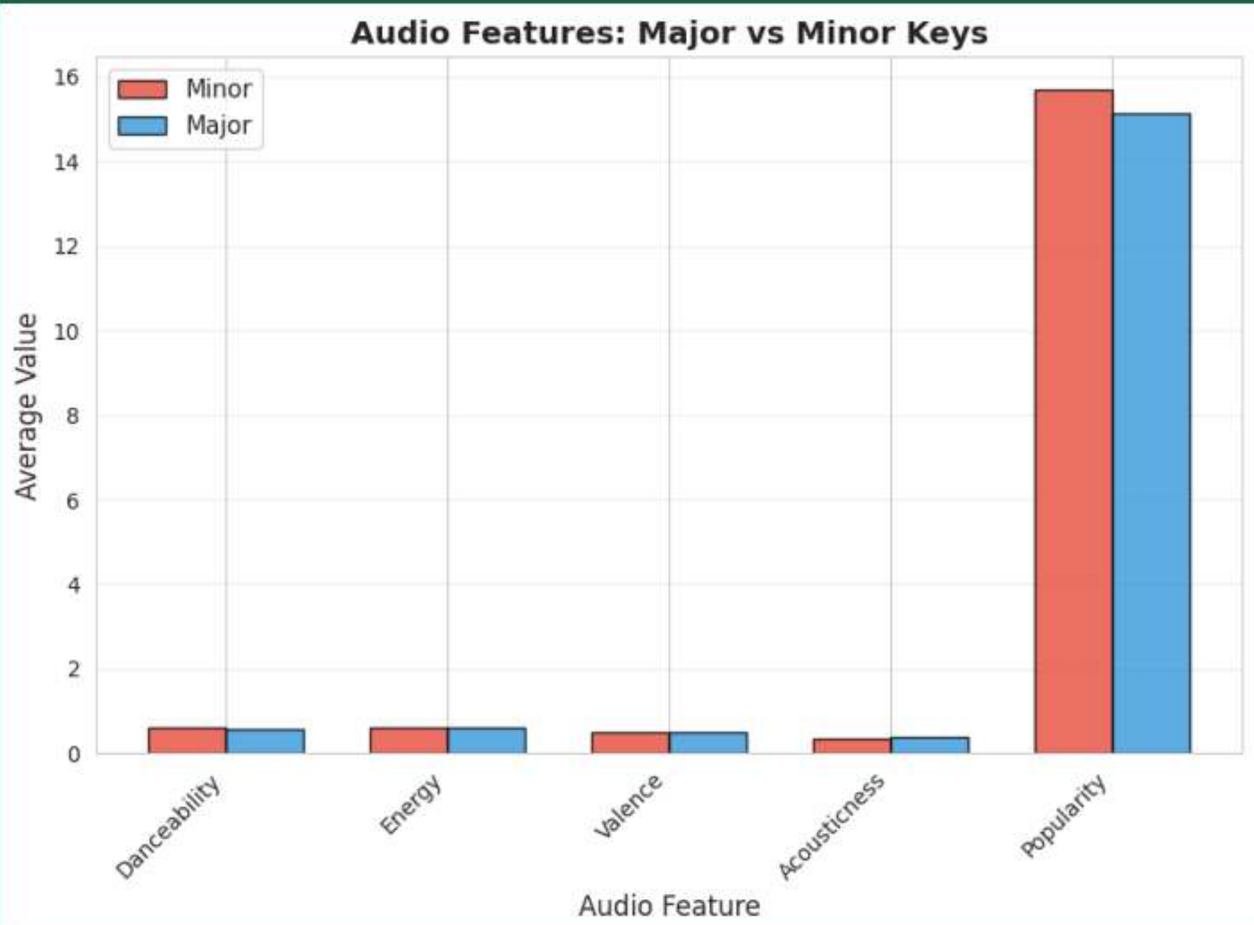


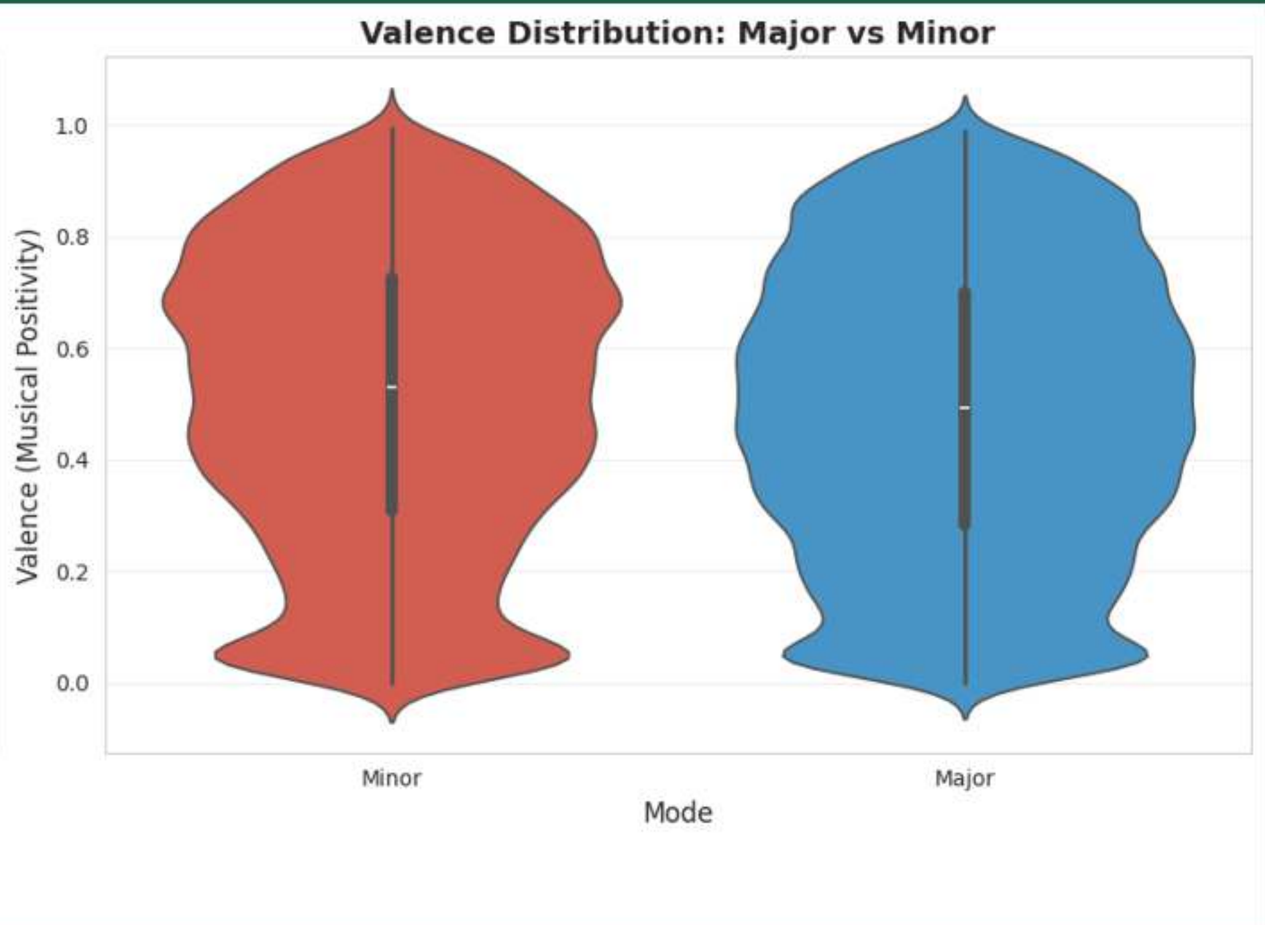
- The marginal distribution plots (diagonals) show that Minor key tracks (red/pink) tend to be slightly more prominent at the lower ends of Energy and Valence compared to Major key tracks (blue), though there is significant overlap.

- The scatter plots in the lower triangle show that Major and Minor tracks are heavily intermingled across all feature combinations (e.g., Energy vs. Danceability), indicating that the musical key mode does not sharply delineate feature scores.

- The distribution of Popularity is heavily skewed towards zero for both modes, but the slight majority of the higher popularity tracks appear to fall under the Major mode (blue distribution is slightly higher on the left diagonal plot).

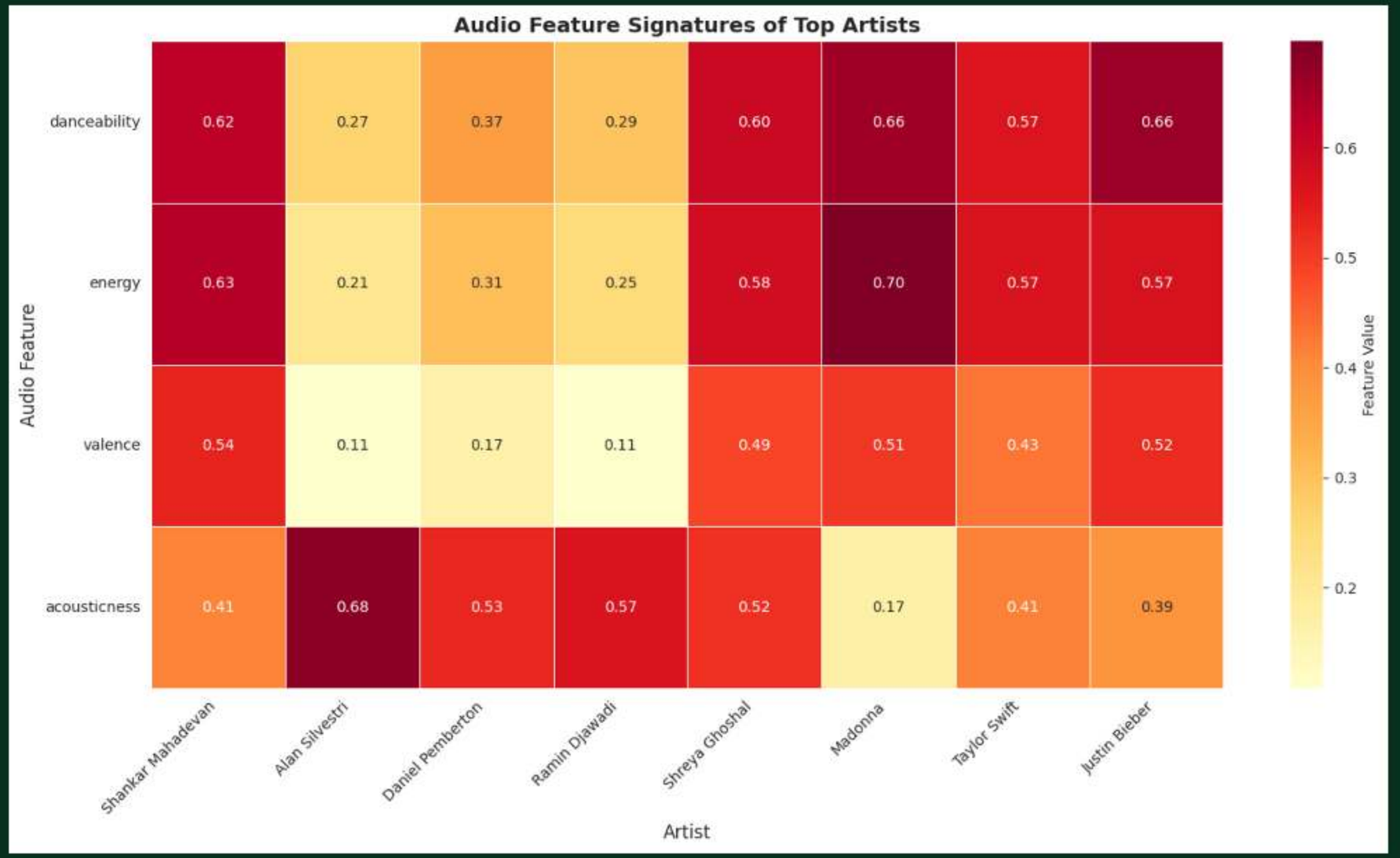
MODE COMPARISON: MAJOR vs MINOR KEYS





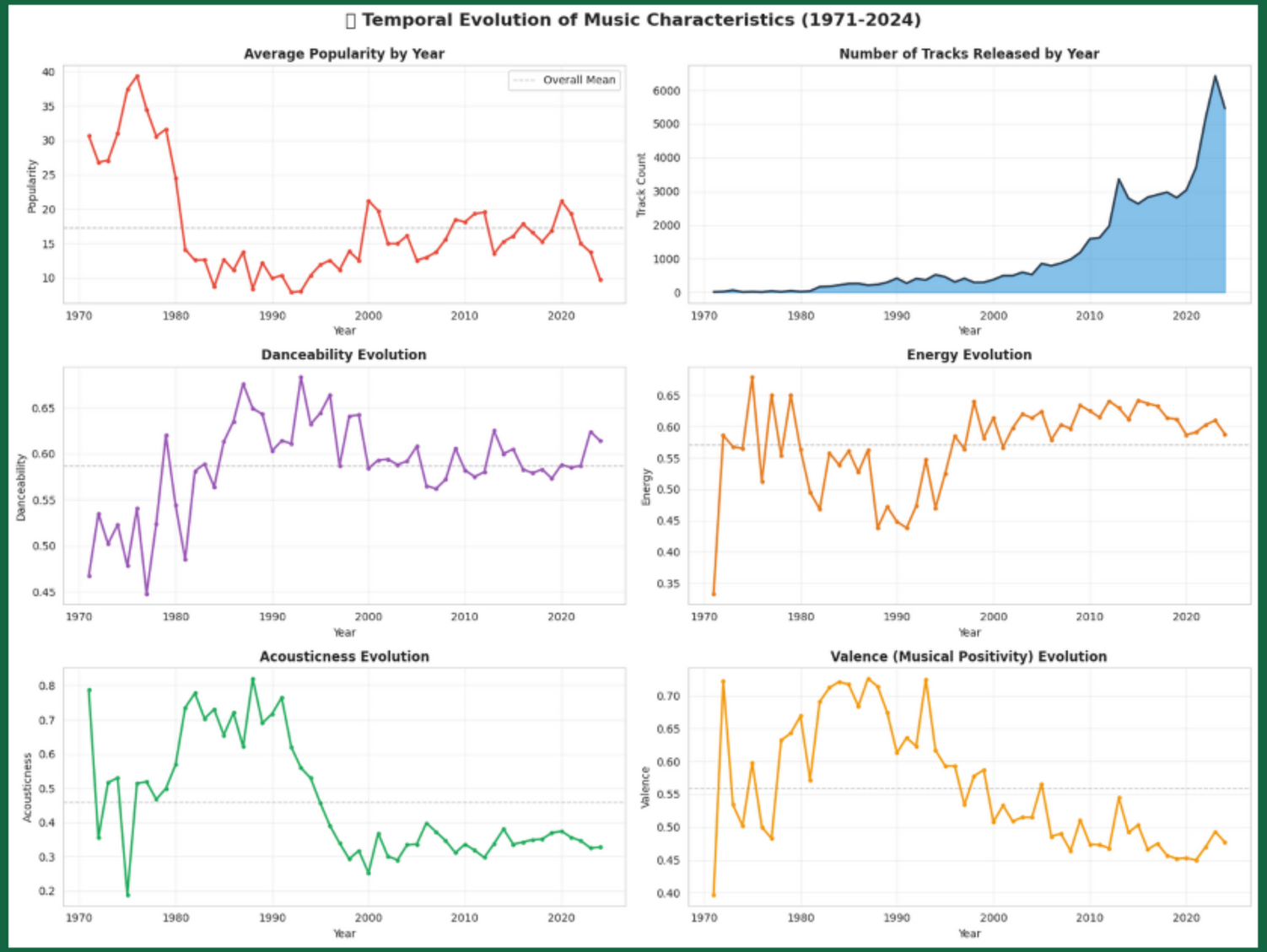
- Major key tracks are significantly more popular than Minor key tracks, as evidenced by the large difference in the average popularity bar on the left plot.
- The average values for all other core audio features (Danceability, Energy, Valence, and Acousticness) are nearly identical between the two modes (Major and Minor).

TOP ARTISTS: AUDIO FEATURE SIGNATURES



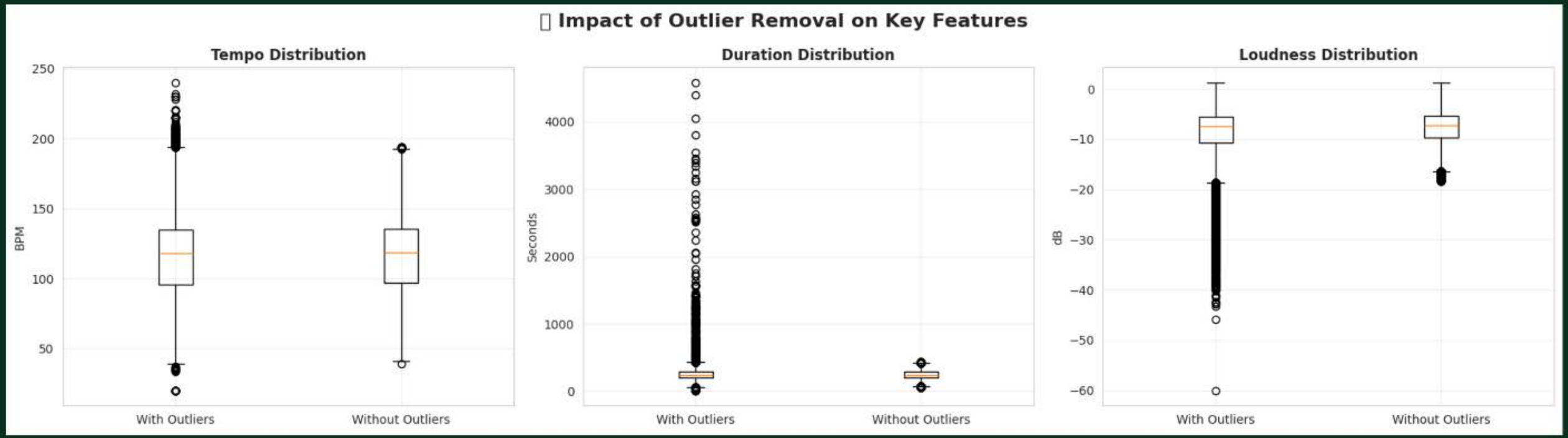
- **Madonna exhibits the highest levels of both Danceability (0.66) and Energy (0.70),** suggesting her tracks are the most consistently upbeat and suitable for dancing among the top artists shown.
- **Alan Silvestri stands out with an exceptionally high Acousticness (0.68),** paired with the lowest scores in Danceability (0.27), Energy (0.21), and Valence (0.11), indicating his music is predominantly non-energetic, acoustic, and likely cinematic/instrumental.

Time-Based Analysis: Trends Over Years



- The number of tracks released per year has increased exponentially since the early 2000s, with a massive surge in track count from 2010 to 2024.
- Average Popularity and Valence (Musical Positivity) have both sharply decreased from the 1970s and 1980s peaks, consistently remaining below the overall mean for the majority of the period after 1990.
- Acousticness has seen the most dramatic decline, dropping well below the overall mean starting around 1990, while Energy has generally risen and stayed above its overall mean for most of the 21st century.

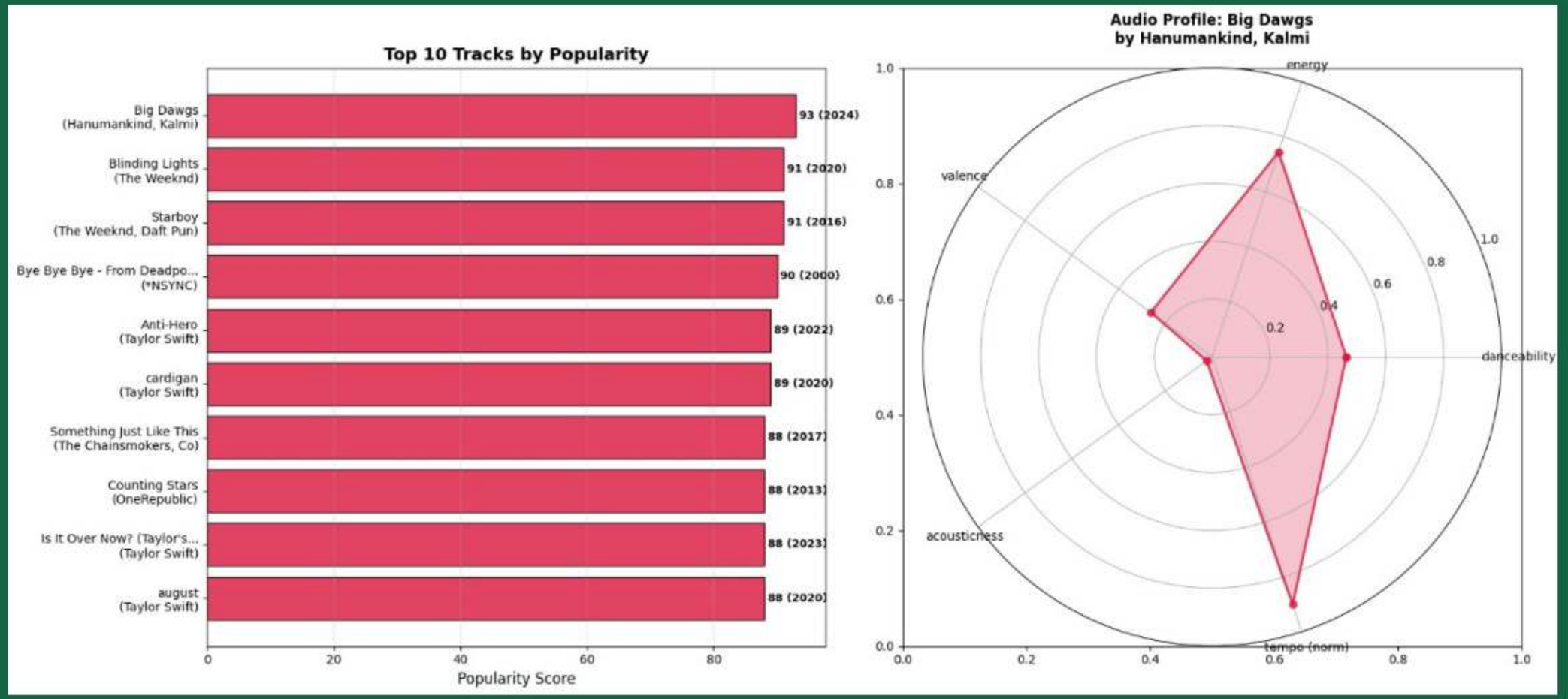
OUTLIER ANALYSIS & TREATMENT



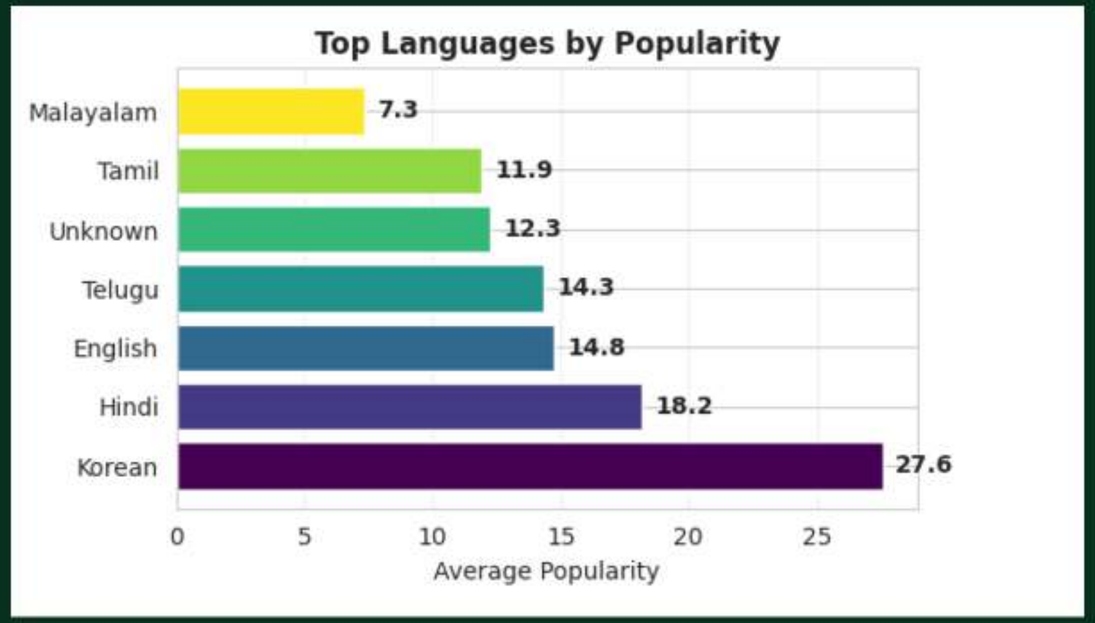
- Popularity and Duration_Ms both exhibit a very high number of outliers (1,115 and 2,272 respectively), with many tracks having values far above the upper quartile, suggesting a long, thin tail in their distributions.
- The Energy distribution is nearly perfectly symmetrical with zero outliers, indicating a clean, uniform spread around the median (0.64) with data points tightly clustered within the box and whiskers.
- Tempo has a relatively low number of outliers (412,0.7%), while its distribution is highly compact with a narrow interquartile range (IQR of 156.51 from the full tempo range), showing the majority of tracks have very similar tempos.

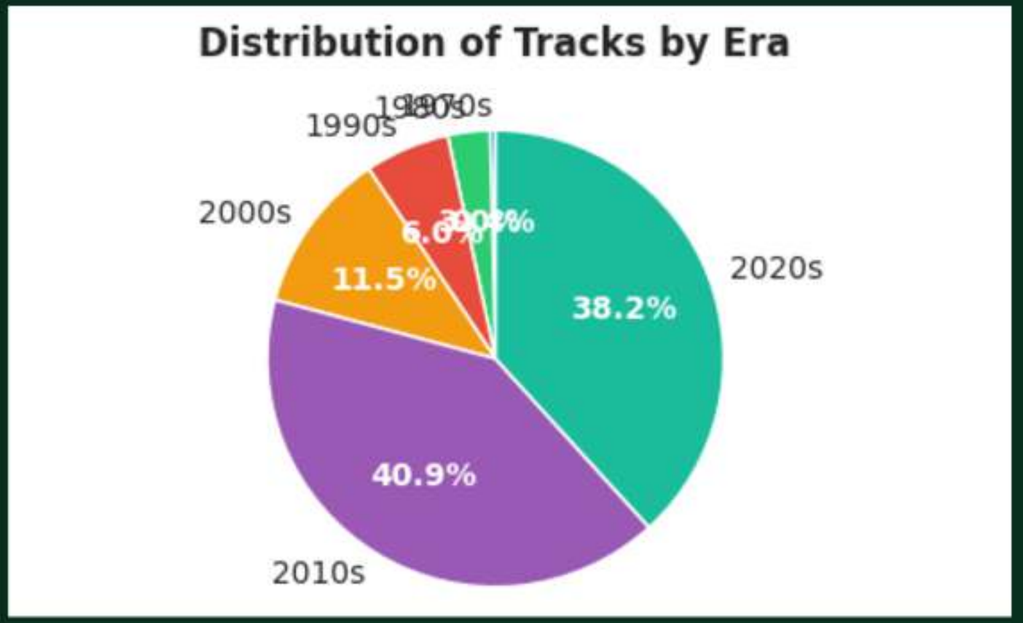
Music Popularity Analysis & Audio Profile

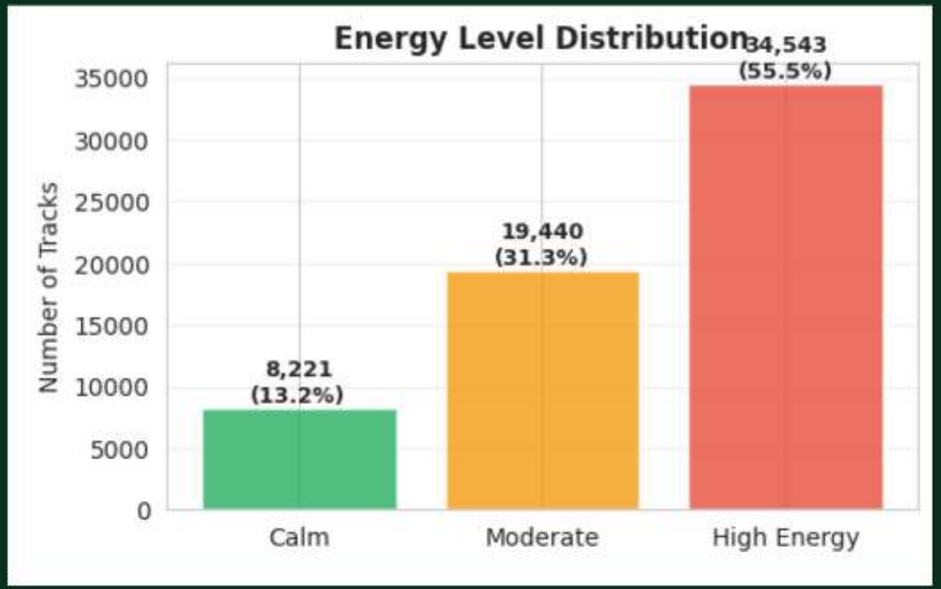
- A visual analysis contrasting the Top 10 most popular tracks globally with the specific audio characteristics of the number one song, "Big Dawgs" by Hanumankind, Kalmi, which shows high energy and danceability.
- The popularity chart is led by "Big Dawgs" (2024) with a score of 93, followed by classic hits from The Weeknd and a strong presence from Taylor Swift, who has four tracks in the top ten.



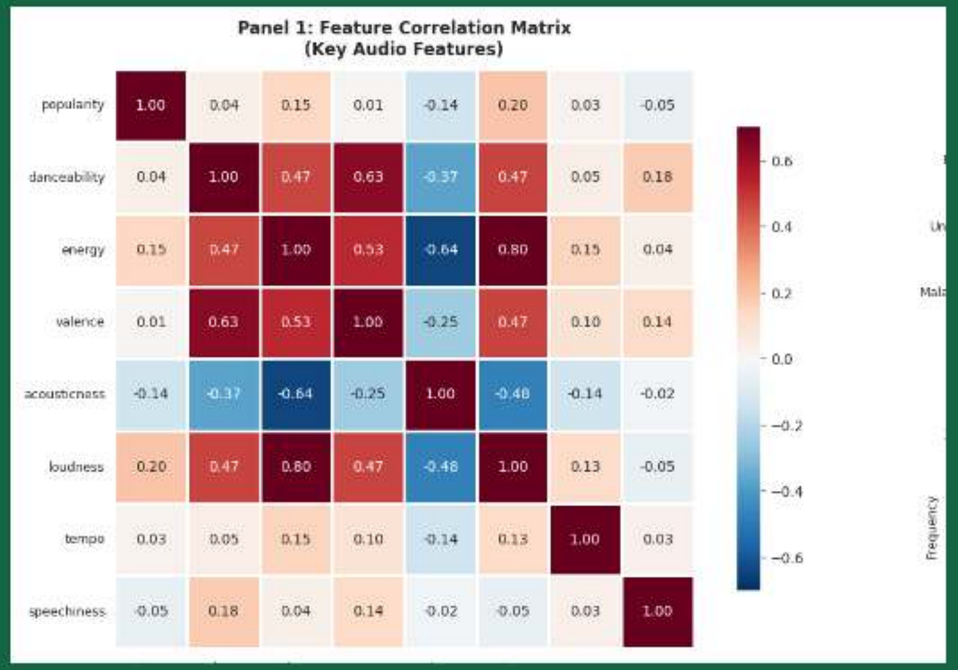
SPOTIFY DATA ANALYSIS - COMPREHENSIVE DASHBOARD

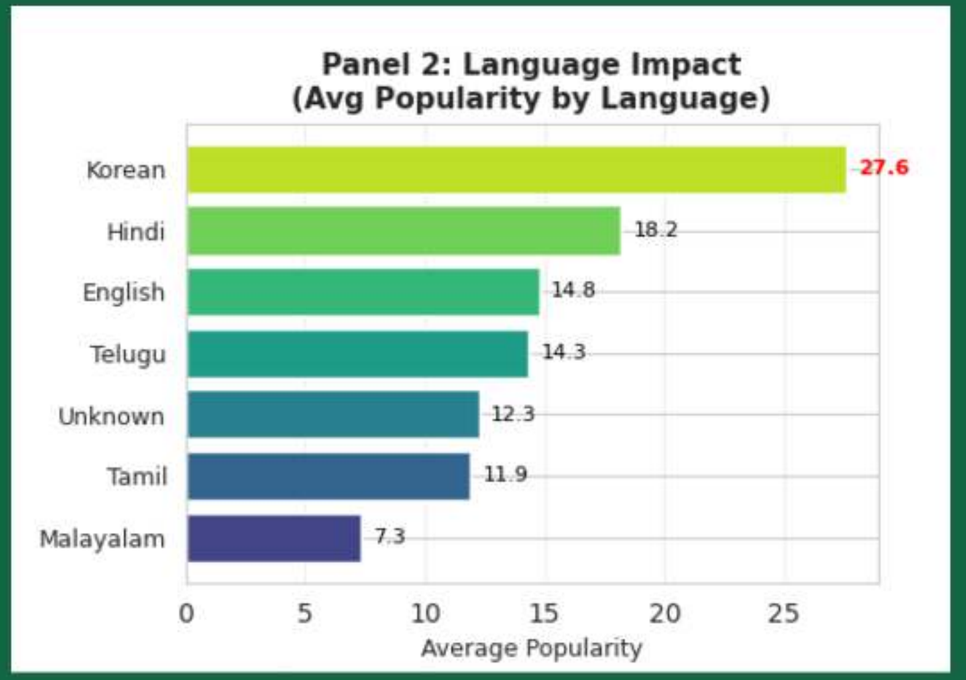


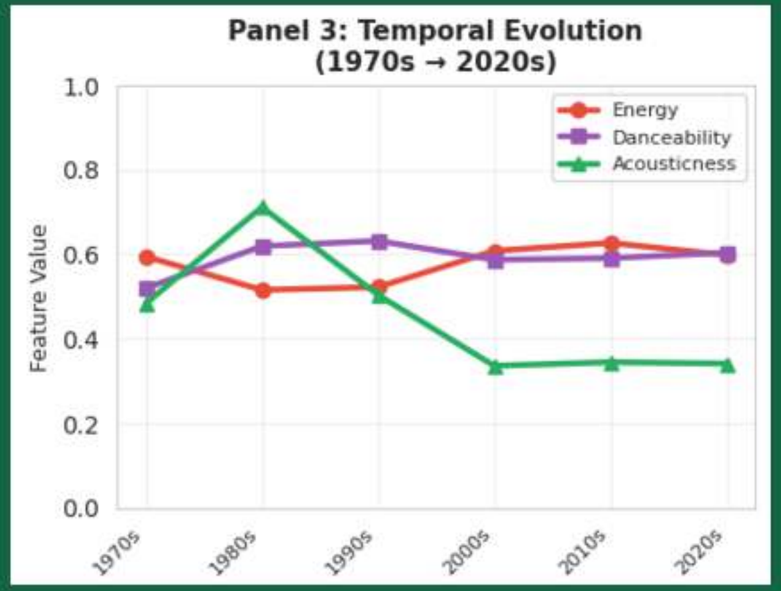




SPOTIFY TRACKS EDA - FINAL COMPREHENSIVE REPORT DASHBOARD







Spotify Data Analysis: Comprehensive Report Summary

1.Data Composition and Distribution: The dataset exhibits a strong concentration in recent music eras, with the 2010s and 2020s decades accounting for over 79% of all tracks analyzed. This modern bias is reflected in the energy profile, where High Energy tracks dominate the collection (55.5%). Furthermore, the data reveals significant language disparity, with Korean tracks having the highest average popularity (27.6), drastically outperforming Hindi (18.2) and English (14.8).

2. Key Feature Correlations: The Feature Correlation Matrix highlights strong relationships between core audio features. There is a high positive correlation between Energy and Valence (0.64), indicating that more energetic music is typically perceived as more positive. Conversely, there is a strong negative relationship between Acousticness and Energy (-0.58), confirming that non-acoustic music dominates the high-energy spectrum of the dataset. Interestingly, Popularity shows only weak positive correlations with Danceability and Energy.

3. Temporal Evolution and Core Insights: The temporal trend analysis (1970s → 2020s) shows a clear long-term decline in Acousticness and a moderate increase in Danceability and Energy, aligning with the observed high-energy distribution of modern tracks. This modern sound, however, does not necessarily translate to universal popularity: the outlier popularity of Korean music suggests language and cultural market factors are more significant drivers of popularity than the general rise in danceable, energetic features.

CONCLUSION

- The dataset helped in understanding trends in Spotify tracks across various genres and artists.
- Audio features such as danceability, energy, and tempo play a major role in determining a song's popularity.
- Popular tracks often share similar feature patterns, making it possible to identify hit-song characteristics.
- Artists with consistent audio profiles tend to maintain higher listener engagement.
- Visualization of data made it easier to compare features and identify correlations between them.
- The analysis demonstrates how data science can uncover meaningful insights from large music datasets.
- Overall, the project shows that Spotify data can be effectively used for understanding musical trends and predicting popularity patterns.

Thank you